SCIENTIFIC OPINION

EFSA Journal

ADOPTED: 12 July 2017

doi: 10.2903/j.efsa.2017.4971

Guidance on the use of the weight of evidence approach in scientific assessments

EFSA Scientific Committee,
Anthony Hardy, Diane Benford, Thorhallur Halldorsson, Michael John Jeger, Helle Katrine Knutsen, Simon More, Hanspeter Naegeli, Hubert Noteborn, Colin Ockleford, Antonia Ricci, Guido Rychen, Josef R Schlatter, Vittorio Silano, Roland Solecki, Dominique Turck, Emilio Benfenati, Qasim Mohammad Chaudhry, Peter Craig, Geoff Frampton, Matthias Greiner, Andrew Hart, Christer Hogstrand, Claude Lambre, Robert Luttik, David Makowski, Alfonso Siani, Helene Wahlstroem, Jaime Aguilera, Jean-Lou Dorne, Antonio Fernandez Dumont, Michaela Hempen, Silvia Valtueña Martínez, Laura Martino, Camilla Smeraldi, Andrea Terron, Nikolaos Georgiadis and Maged Younes

Abstract

EFSA requested the Scientific Committee to develop a guidance document on the use of the weight of evidence approach in scientific assessments for use in all areas under EFSA's remit. The guidance document addresses the use of weight of evidence approaches in scientific assessments using both qualitative and quantitative approaches. Several case studies covering the various areas under EFSA's remit are annexed to the guidance document to illustrate the applicability of the proposed approach. Weight of evidence assessment is defined in this guidance as a process in which evidence is integrated to determine the relative support for possible answers to a question. This document considers the weight of evidence assessment as comprising three basic steps: (1) assembling the evidence into lines of evidence of similar type, (2) weighing the evidence, (3) integrating the evidence. The present document identifies reliability, relevance and consistency as three basic considerations for weighing evidence.

Keywords: risk assessment, weight of evidence, biological relevance, uncertainty, lines of evidence

Requestor: EFSA Scientific Committee

Question number: EFSA-Q-2015-00007

Correspondence: sc.secretariat@efsa.europa.eu

Scientific Committee members: Anthony Hardy, Diane Benford, Thorhallur Halldorsson, Michael John Jeger,Helle Katrine Knutsen, Simon More, Hanspeter Naegeli, Hubert Noteborn, Colin Ockleford, Antonia Ricci, Guido Rychen, Josef R Schlatter, Vittorio Silano, Roland Solecki, Dominique Turck and Maged Younes.

Acknowledgements: The Panel wishes to thank the members of the Working Group on the Weight of Evidence Approach in Scientific Assessments: Emilio Benfenati, Qasim Mohammad Chaudhry, Peter Craig, Geoff Frampton, Matthias Greiner, Andrew Hart, Christer Hogstrand, Claude Lambre, Robert Luttik, David Makowski, Alfonso Siani, Maged Younes (Chair) and Helene Wahlstroem for the preparatory work on this scientific output, George Fotakis (ECHA observer) and Kristina Thayer (Hearing Expert) for their input and EFSA staff members: Jaime Aguilera, Jean-Lou Dorne, Antonio Fernandez Dumont, Michaela Hempen, Silvia Valtueña Martínez, Laura Martino, Camilla Smeraldi, Andrea Terron and Nikolaos Georgiadis for the support provided to this scientific output.

Suggested citation: EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Benfenati E, Chaudhry QM, Craig P, Frampton G, Greiner M, Hart A, Hogstrand C, Lambre C, Luttik R, Makowski D, Siani A, Wahlstroem H, Aguilera J, Dorne J-L, Fernandez Dumont A, Hempen M, Valtueña Martínez S, Martino L, Smeraldi C, Terron A, Georgiadis N and Younes M, 2017. Scientific Opinion on the guidance on the use of the weight of evidence approach in scientific assessments. EFSA Journal 2017;15(8):4971, 69 pp. https://doi.org/10.2903/j.efsa.2017.4971

ISSN: 1831-4732

The EFSA Journal is a publication of the European Food Safety Authority, an agency of the European Union.

Summary

The European Safety Authority (EFSA) requested the Scientific Committee (SC) to develop a guidance document on the use of the weight of evidence approach in scientific assessments for use in all areas under EFSA's remit.

The guidance document addresses the use of weight of evidence approaches in scientific assessments using both qualitative and quantitative approaches. Several case studies covering the various areas under EFSA's remit are annexed to the guidance document to illustrate the applicability of the proposed approach.

In developing the guidance, the Working Group (WG) of the SC took into account other EFSA activities and related European and international activities to ensure consistency and harmonisation of methodologies in order to provide an international dimension to the guidance and avoid duplication of the work.

This guidance document is intended to guide EFSA Panels and staff on the use of the weight of evidence approach in scientific assessments. It provides a flexible framework that is applicable to all areas within EFSA's remit, within which assessors can apply those methods which most appropriately fit the purpose of their individual assessment.

Weight of evidence assessment is defined in this guidance as a process in which evidence is integrated to determine the relative support for possible answers to a question. This document considers the weight of evidence assessment as comprising three basic steps: (1) assembling the evidence into lines of evidence of similar type, (2) weighing the evidence, (3) integrating the evidence.

The present document identifies reliability, relevance and consistency as three basic considerations for weighing evidence. They are defined in terms of their contributions to the weight of evidence assessment: Reliability is the extent to which the information comprising a piece or line of evidence is correct. Relevance is the contribution a piece or line of evidence would make to answer a specified question, if the information comprising the line of evidence was fully reliable. Consistency is the extent to which the contributions of different pieces or lines of evidence to answering the specified question are compatible.

While no specific methods are prescribed, a list of criteria for comparing weight of evidence methods is provided to assist in evaluating the relative strengths and weaknesses of the different methods. The criteria do not necessarily have equal importance: their relative importance may be considered on a case-by-case basis when planning each weight of evidence assessment.

All EFSA scientific assessments must include consideration of uncertainties, reporting clearly and unambiguously what sources of uncertainty have been identified and what their impact on the assessment outcome is.

Reporting should be consistent with EFSA's general principles regarding transparency and reporting. In a weight of evidence assessment, this should include justifying the choice of methods used, documenting all steps of the procedure in sufficient detail for them to be repeated, and making clear where and how expert judgement has been used. Reporting should also include referencing and, if appropriate, listing or summarising all evidence considered, identifying any evidence that was excluded; detailed reporting of the conclusions; and sufficient information on intermediate results for readers to understand how the conclusions were reached.

EFSA Panels and Units are encouraged to review their existing approaches to weight of evidence assessment in the light of the guidance document, and to consider in particular:

- Whether all pertinent aspects of reliability, relevance and consistency are addressed,
- How to ensure the transparency of weight of evidence assessments,
- Carry out some case studies to assess whether additional methods described in the guidance would add value to their scientific assessments.

Table of contents

1. Introduction

1.1. Background and Terms of Reference as provided by EFSA

EFSA's Science Strategy 2012–2016 has identified four strategic objectives: (1) further develop excellence of EFSA's scientific advice, (2) optimise the use of risk assessment capacity in the European Union (EU), (3) develop and harmonise methodologies and approaches to assess risks associated with the food chain, (4) strengthen the scientific evidence for risk assessment and risk monitoring. In this context, the harmonisation and development of new methodologies for risk assessment and scientific assessments is of critical importance to deliver EFSA's science strategy. For this purpose, a number of projects have recently started at the European Food Safety Authority (EFSA) to address individual and cross-cutting methodological issues within the whole scientific assessment landscape. The Assessment Methodological unit (AMU) of EFSA has started a project (PROMETHEUS) with the objective of supporting the coordination and consistency of all EFSA projects that aim at developing or refining methodological approaches. Such an umbrella project will provide a definition of the guiding principles for evidence-based assessments and a collection of available approaches and will identify areas where methods or tools are needed to fulfil such guiding principles.[1]

In July 2013, the Scientific Committee (SC) of EFSA published an opinion on 'priority topics for the development of risk assessment guidance by EFSA's SC' which used a number of criteria to make recommendations for the preparation of new or the revision of existing guidance documents as follows:

- Across Panel relevance
- Critical importance including urgency of topic to be addressed for several Panels
- Topic not being addressed by an individual Panel
- Sufficient information available to develop meaningful guidance
- International dimension.

From this prioritisation exercise, the SC opinion identified three priority topics for 2014: uncertainty analysis, biological relevance, and the use of the weight of evidence (weight of evidence) in scientific assessments (EFSA Scientific Committee, 2013).

The latter is the subject of this project. The weight of evidence has been defined by the WHO as 'a process in which all of the evidence considered relevant for a risk assessment is evaluated and weighted' (WHO, 2009). The SC of EFSA used the WHO definition and pointed out that evidence can be derived from several sources such as white literature (peer reviewed scientific publications), grey literature (reports on websites of governmental, nongovernmental, intragovernmental agencies, etc.) and black literature (confidential reports). In order to increase transparency in the risk and other scientific assessment processes, it is important to provide a methodology to select, weigh and integrate the evidence in a systematic, consistent and transparent way to reach the final conclusions and to identify related uncertainties (SCENIHR, 2012; EFSA Scientific Committee, 2013). In addition, the SC of EFSA noted that part of the overall weighing of the evidence deals with the evaluation of equivalent or similar questions performed by other international bodies and the adequacy of such evaluations should be judged by EFSA before taking them into account. This is particularly helpful in cases for which the information available is so extensive that it is beyond the capability of a single evaluation to judge each individual study, report, publication by itself. In addition, systematic reviews (SRs) may be very useful. However, the adequacy of the process, the pertinence to the risk assessment, the nature of the question and the inclusion and exclusion criteria should be transparently evaluated by EFSA before taking SRs into account (EFSA Scientific Committee, 2013). Considering the example of chemical risk assessment, the weight of evidence approach requires expert judgement of distinct lines of evidence (*in vivo*, *in vitro*, *in silico*, population studies, modelled and measured exposure data, etc.), which may come from studies conducted according to official guidelines (e.g. OECD) or from non-standardised methodologies. In this context, data from all sources and categories of literature should be considered for the risk assessment processes, as appropriate to determine their quality and relevance. These considerations should then be reflected in the relative weight given to the evidence in the scientific assessment and transparently taken into account in the overall evaluation of uncertainty (EFSA Scientific Committee, 2013). It is therefore proposed that the SC of EFSA develop guidance on the use of the weight of evidence approach in scientific assessments.

[1] For current information on PROMETHEUS see: https://www.efsa.europa.eu/en/methodology/evidence

Terms of Reference

EFSA requests the SC to develop a guidance document on the use of the weight of evidence approach in scientific assessments for use in all areas under EFSA's remit.

The guidance document should address the use of the weight of evidence in scientific assessments using both qualitative and quantitative approaches. Several case studies covering the various areas under EFSA's remit should be annexed in the guidance document to illustrate the proposed approaches.

In developing the guidance, the Working Group (WG) of the SC should take into account other EFSA activities and related European and international activities to ensure consistency and harmonisation of methodologies, to provide an international dimension to the guidance and avoid duplication of the work.

In line with EFSA's policy on openness and transparency, EFSA will publish a draft version of the guidance document for public consultation to invite comments from the scientific community and stakeholders. Subsequently, the guidance document and the results of the public consultation should be presented at an international event after publication.

1.2. Interpretation of the Terms of Reference

In the context of risk assessment, various formal definitions and synonyms have been offered by IPCS (2004), US EPA (2000), WHO FAO, US National Research Council's Committee, SCHER, SCENIHR, SCCS (2013) on Improving Risk Analysis Approaches for the phrase 'weight of evidence' or 'evidence synthesis'.

When addressing the mandate, the SC acknowledged that the issue of weight of evidence approaches in risk assessment encompasses aspects related to the reliability of the various pieces of evidence used in the assessment.

In order for the guidance document to address the use of the weight of evidence approaches in scientific assessments using both qualitative and quantitative approaches, a list of the available approaches used globally has been provided together with several case studies from various areas under EFSA's remit to illustrate the proposed approaches.

In developing the guidance, the WG of the SC has taken into account other EFSA activities and related European and international activities to ensure consistency and harmonisation of methodologies, to provide an international dimension to the guidance and avoid duplication of the work.

In particular:

- relevant guidance published by the SC on related subjects (transparency in risk assessment, uncertainty in exposure assessment, statistical significance and biological relevance (EFSA, 2006, 2009, 2011) and the latest draft guidance documents on uncertainty and biological relevance that are being developed concomitantly.
- the guidance on uncertainty analysis in risk assessment. It deals specifically with reporting and analysing uncertainties using qualitative and quantitative methods for all work within EFSA's remit. The overlaps with the weight of evidence approach should be carefully taken into account by both WGs to ensure that the weight of evidence and uncertainty guidance documents are consistent, use harmonised methodologies, and do not duplicate the work.
- the guidance on biological relevance for all areas of work within EFSA's remit. It deals specifically with criteria to evaluate biological relevance in scientific assessments and the overlaps with the weight of evidence approach should be carefully taken into account by both WGs to ensure consistency, the use of harmonised methodologies, and to avoid duplication.
- the current work within EFSA on the approach to evidence-based risk assessment to ensure consistency, the use of harmonised methodologies, and to avoid duplication of the work.

The guidance was also expected to take into account other European and international activities:

- The work of the European Commission's non-food scientific committees and other agencies on weight of evidence approach, and where appropriate, seek their participation for the development of a harmonised guidance.

Examples include best practices in weight of evidence methodologies, the use of SR in risk assessment, the WHO application of the weight of evidence approach in relation to the mode of action framework and other related international developments (EFSA, 2010a; Meek et al., 2014; Perkins et al., 2015; OECD, 2016; Wittwehr et al., 2016).

1.3. Aim and scope of the document

Weighing the evidence is an inherent part of every scientific assessment performed by EFSA. Experts review all available data, and come to conclusions based on an assessment of their overall confidence in the results of all reviewed studies. The approaches and methods used in conducting such an 'non-formalised', inherent weighing of the evidence are mostly not spelled out, however.

The aim of this guidance document is to provide a general framework for considering and documenting the approaches used to weigh the evidence in answering the main question of each scientific assessment or questions that need to be answered in order to provide, in conjunction, an overall answer. The document further indicates, in general terms, types of qualitative and quantitative approaches to weigh and integrate evidence, and lists individual methodologies with pointers as to where details of these can be found. Finally, the document provides suggestions for conducting and reporting of weight of evidence assessments.

The document does not attempt to prescribe approaches or methods to be used, nor does it provide a comprehensive description of all methods that can be used.

1.4. Relation to other relevant EFSA guidance documents

The guidance on the use of the weight of evidence approaches builds on the conceptual approach for scientific assessments as described in PROMETHEUS (EFSA, 2015b), which describes the overall process for dealing with data and evidence.

Transparent reporting of all assumptions and methods used, including expert judgement, is necessary to ensure that the assessment process leading to the conclusions is fully comprehensible.

'Open EFSA' aspires both to improve the overall quality of the available information and data used for its scientific outputs and to comply with normative and societal expectations of openness and transparency (EFSA, 2009, 2014a–d). In line with this, EFSA is publishing three separate but closely related guidance documents to guide its expert Panels for use in their scientific assessments (EFSA, 2015a–c). These documents address three key elements of the scientific assessment: the analyses of Uncertainty, Weight of Evidence assessment and Biological Relevance.

The first document provides guidance on how to identify, characterise, document and explain all types of uncertainty arising within an individual assessment for all areas of EFSA's remit. The Guidance does not prescribe which specific methods should be used from the toolbox but rather provides a harmonised and flexible framework within which different described qualitative and quantitative methods may be selected according to the needs of each assessment.

This current document on weight of evidence assessment provides a general framework for considering and documenting the approach used to evaluate and weigh the assembled evidence when answering the main question of each scientific assessment or questions that need to be answered in order to provide, in conjunction, an overall answer. This includes assessing the relevance, reliability and consistency of the evidence. The document further indicates the types of qualitative and quantitative methods that can be used to weigh and integrate evidence and points to where details of the listed individual methods can be found. The weight of evidence approach carries elements of uncertainty analysis that part of uncertainty which is addressed by weight of evidence analysis does not need to be reanalysed in the overall uncertainty analysis, but may be added to.

The third document provides a general framework to addresses the question of biological relevance at various stages of the assessment: the collection, identification and appraisal of relevant data for the specific assessment question to be answered. It identifies generic issues related to biological relevance in the appraisal of pieces of evidence, in particular, and specific criteria to consider when deciding on whether or not an observed effect is biologically relevant, i.e. whether it shows an adverse or a positive health effect. A decision tree is developed to aid the collection, identification and appraisal of relevant data for the specific assessment question to be answered. The reliability of the various pieces of evidence used and how they should be integrated with other pieces of evidence is considered by the weight of evidence guidance document.

EFSA will continue to strengthen links between the three distinct but related topics to ensure the transparency and consistency of its various scientific outputs while keeping them fit for purpose.

1.5. Audience and degree of obligation

This Guidance is aimed at all those contributing to EFSA assessments and provides a harmonised, but flexible framework that is applicable to all areas of EFSA's work and all types of scientific assessment, including risk assessment. In line with improving transparency, the SC considers the application of this guidance to be unconditional for EFSA. Each assessment must clearly and unambiguously document:

- what evidence was considered and how it was assembled into lines of evidence;
- how the evidence was weighed and integrated including consideration of reliability, relevance and consistency;
- the conclusion on the weight of evidence question in terms of the range and probability of possible answers. This can be expressed qualitatively or quantitatively, but should be quantified if possible when it directly addresses the Terms of Reference for the assessment.

The document provides guidance on the general principles of the weight of evidence approach but assessors have the flexibility to choose appropriate methods, and the degree of refinement in applying them.

2. General framework and principles for weight of evidence assessment

This section provides a general framework and principles for weight of evidence assessment, including definitions of key concepts. Many scientific assessments involve weighing of evidence, although this may be implicit rather than explicit and is only sometimes described as 'weight of evidence assessment'. The aim of this guidance is to make weight of evidence assessment more explicit and transparent, and to provide a general framework of principles and approaches which is applicable to all areas of EFSA's work. Account is taken of approaches already used by EFSA, by other EU and international organisations, and in the scientific literature.

2.1. Weight of evidence assessment and lines of evidence

WHO (2009) has defined weight of evidence assessment as 'a process in which all of the evidence considered relevant for a risk assessment is evaluated and weighted'. A recent review by ANSES (2016) defines weight of evidence assessment as 'the structured synthesis of lines of evidence, possibly of varying quality, to determine the extent of support for hypotheses'. Definitions and descriptions from a selection of other relevant publications are presented in Annex B, and reflect similar concepts to those of WHO and ANSES. The core of most definitions is that weight of evidence assessment is a process for integrating evidence to arrive at conclusions.

In practice, weighing[2] of evidence may occur when estimating quantities, as well as when assessing hypotheses, and both are relevant to EFSA's work. Therefore, this document uses a broader definition, as follows:

Weight of evidence assessment is a process in which evidence is integrated to determine the relative support for possible answers to a scientific question.

The term 'weight of evidence' on its own is the extent to which evidence supports possible answers to a scientific question. This is what is assessed by weight of evidence assessment, and can be expressed qualitatively or quantitatively (discussed further in Section 2.3, below).

It is often useful to organise evidence into groups or categories, which are often referred to as lines of evidence. ANSES (2016) defines a line of evidence as 'a set of relevant information of similar type grouped to assess a hypothesis'. Rooney et al. (2014) use the variant 'streams of evidence' which they describe as referring specifically to human data, animal data, and 'other relevant data (including mechanistic or *in vitro* studies)'. This document simplifies the ANSES (2016) definition, replacing 'relevant information' with 'evidence' and reducing the emphasis on hypotheses, because weight of evidence assessment may be applied to quantities as well as hypotheses, as mentioned above. This results in the following general definition, which is compatible with other uses of the same term in the literature (see Annex B):

[2] The term 'weighing' is used in preference to 'weighting' in this document, as not all approaches to weighing evidence involve explicit assignment of 'weights'.

A **line of evidence** is a set of evidence of similar type.

Various terms have been used to refer to distinct elements of information within a line of evidence, including 'studies' and 'pieces of evidence'. **Piece of evidence** is a more general term, as it could refer to a study (or to one of multiple outcomes of a study), or to other types of information including expert knowledge, experience, a model or even a single observation. In some cases, a line of evidence may comprise only a single piece of evidence.

Pieces of evidence may show varying degrees of similarity. There is no fixed rule on how much similarity is required within the same line of evidence. This is for the assessor(s) to decide, and depends on what they find useful for the purpose of the scientific assessment. For example, in some assessments, it might be sufficient to treat all human studies as a single line of evidence, whereas in other assessments it might be helpful to treat different types of human studies as separate lines of evidence.

The definition for line of evidence is broadly worded to accommodate different ways in which lines of evidence may contribute to answering a question. Different lines of evidence for the same question may be standalone, in the sense that each line of evidence offers an answer to the question without needing to be combined with other lines of evidence. It is important to distinguish these from complementary lines of evidence, which can only answer the question when they are combined. Multiple experiments measuring the same parameter are examples of standalone lines of evidence, whereas data on hazard and exposure are complementary lines of evidence for risk assessment because both are necessary and must be combined to assess risk. The distinction between complementary and standalone lines of evidence is important because it has practical implications in weight of evidence assessment (see Section 4). Note that a single question may be addressed by a combination of standalone and complementary lines of evidence.

Assessors often refer to 'data gaps', where types or lines of evidence that would have been useful are lacking. These are easier to detect in regulatory assessments, where lists of required data are established in legislation or guidance. Although weight of evidence assessment is described in terms of evaluating available evidence it also takes account of data gaps. This is because, when a particular type of evidence is absent, the contribution it could have made will also be absent. How much this affects the assessment will depend on the extent to which the available evidence can answer the question itself, or substitute for what is missing (e.g. by read-across) (see also Section 4).

2.2. When to use weight of evidence approaches

In general, the purpose of weight of evidence assessment is to answer a scientific question, as implied in the preceding section. EFSA assessments address questions posed by their Terms of Reference. In some cases, a question in the Terms of Reference may be addressed directly, but in other cases, it is beneficial to divide the primary question into two or more subsidiary questions (EFSA, 2015b).

Weighing of evidence is involved, either explicitly or implicitly, wherever more than one piece of evidence is used to answer a question. Weight of evidence assessment is not needed for scientific questions where no integration of evidence is required.

Thus, a single scientific assessment may comprise one or many questions and none, some or all of those questions may require weight of evidence assessment.

Clarifying the questions posed by the Terms of Reference and deciding whether and how to subdivide them, and whether they require weight of evidence assessment, is part of the first stage of scientific assessment, often referred to as problem formulation. This may show that the question is relatively simple and can be addressed directly, by a straightforward assessment. In many assessments, however, questions may need to be subdivided to yield more directly answerable questions. In this manner, a hierarchy or tree of questions may be established. Assessment then starts at the bottom of the hierarchy. The evidence is divided into lines of evidence, as far as is helpful, assessed, weighed and integrated to answer each question at the bottom of the hierarchy. Integration continues upwards through the question hierarchy following similar principles, until full integration is reached to answer the main question defined by the problem formulation.

In some cases, the Terms of Reference for an assessment pose open questions, for example, to review the state of science on a particular topic. These assessments also require weight of evidence assessment approaches, because their conclusions generally derive from weighing and integrating evidence.

2.3. Weight of evidence conclusions

As implied in the definition above, the purpose of weighing evidence is to assess the relative support for possible answers to a scientific question. In some cases, it may be concluded that the evidence supports only one answer, with complete certainty. More usually, multiple answers remain possible, with differing levels of support. In such cases, the conclusion should state the range of answers that remain possible, and not be reduced to a single answer unless a threshold level of support for conclusions has been agreed with decision-makers, because this involves risk management considerations.

When weight of evidence assessment directly addresses the conclusion of a scientific assessment, its output will be part of the response to the Terms of Reference for the assessment. In general, decision-makers need to know the range of possible answers to their questions, and how probable they are, because this may have important implications for decision-making (EFSA, 2016a–c). Furthermore, it is important to express this quantitatively when possible, to avoid the ambiguity of qualitative expression (EFSA, 2012, 2016a–c).

When weight of evidence addresses an intermediate question in a larger assessment, the possible answers and their relative support needs to be taken into account in subsequent steps of the assessment. In these cases, relative support may be expressed either qualitatively or quantitatively, depending on what is convenient for use in the subsequent steps. Qualitative and quantitative approaches are discussed further in Section 3.

2.4. Steps in weight of evidence assessment

This document considers the weight of evidence assessment as comprising three basic steps:

1) Assembling the evidence,
2) Weighing the evidence,
3) Integrating the evidence.

This corresponds to the three basic steps distinguished by Suter and Cormier (2011, their Figure 3; see also Suter, 2016). The first step involves searching for and selecting evidence that is relevant for answering the question in hand, and deciding whether and how to group it into lines of evidence. The second step involves detailed evaluation and weighing of the evidence. In the third step, the evidence is integrated to arrive at conclusions, which involves weighing the relative support for possible answers to the question.

Practical guidance for the three basic steps is provided in Section 4. Relevant considerations to be taken into account in the weighing and integrating steps are discussed in Section 2.5, while qualitative and quantitative methods for assessing those considerations are discussed in Section 3.

The three steps of weight of evidence assessment, described above, may occur at one or more points in the course of a scientific assessment, wherever integration of evidence is required, as illustrated in Figure 1. The question to be addressed by each weight of evidence assessment is defined by problem formulation, which is a preceding step in the scientific assessment as a whole. The output of weight of evidence assessment feeds either directly or indirectly into the overall conclusion of the scientific assessment. Although weight of evidence assessment itself addresses some of the uncertainty affecting the scientific assessment (see below), a separate step of uncertainty analysis is still needed to take account of any other uncertainties affecting the overall assessment. Some assessments will also include a step of sensitivity analysis or influence analysis, to identify which evidence and uncertainties have most influence on the conclusion.

Any part of the overall assessment may be refined iteratively, when necessary, by returning from later steps to earlier steps, depending on which steps it is most useful to refine.

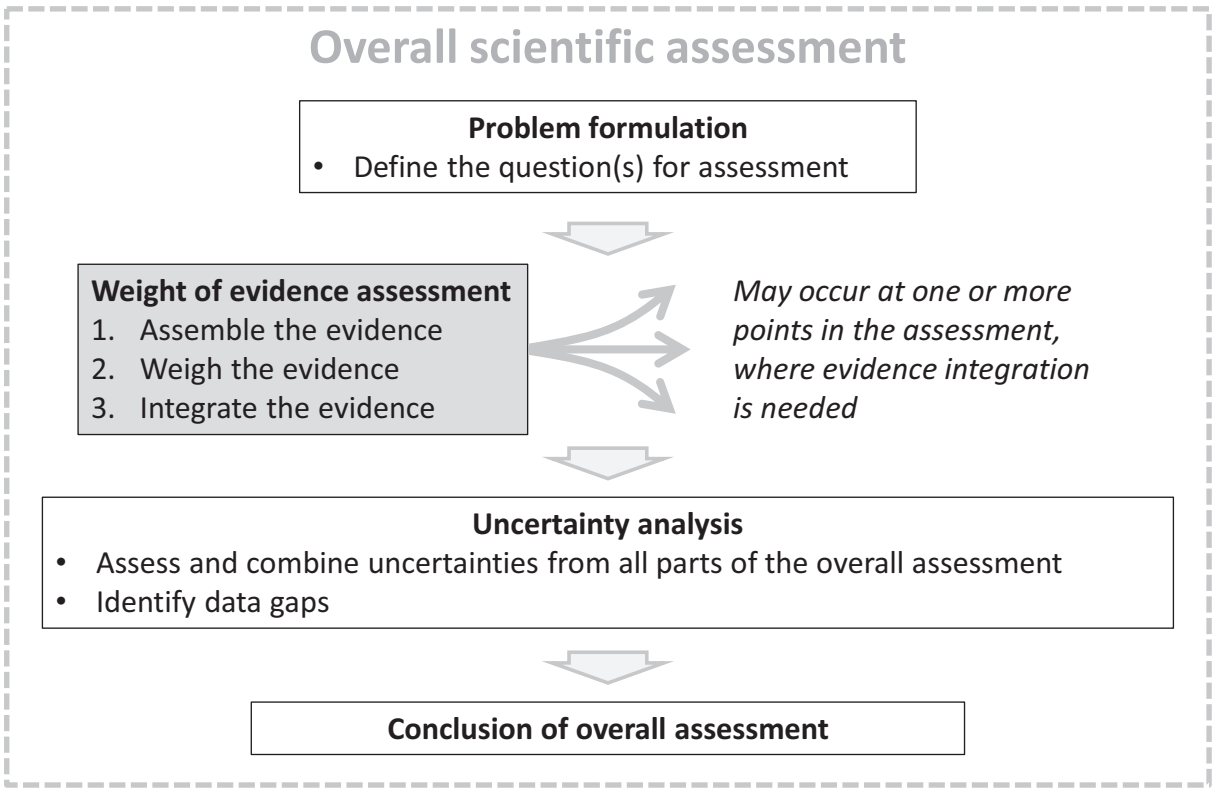

Figure 1: Diagrammatic illustration of weight of evidence assessment as a 3-step process which may occur at one or more points in the course of a scientific assessment

2.5. Key considerations for weighing evidence

Reliability, relevance and consistency are mentioned in many publications on weight of evidence assessment (see Annex B). These can be seen as three basic considerations in the weight of evidence assessment: how applicable the evidence is to the question of interest, the quality of the evidence and how consistent it is with other evidence for the same question. How these three concepts relate to one another, to the three basic steps of weight of evidence assessment and to the weight of evidence conclusion are illustrated in Figure 2. Note that relevance and reliability may be considered in both the first and second steps. First, relevance is considered when identifying evidence, and both relevance and reliability may be considered when selecting which of the identified evidence to include in the assessment (sometimes referred to as screening or filtering). However, the selected evidence will vary in both relevance and reliability and this will be considered in the second step, when weighing the evidence.

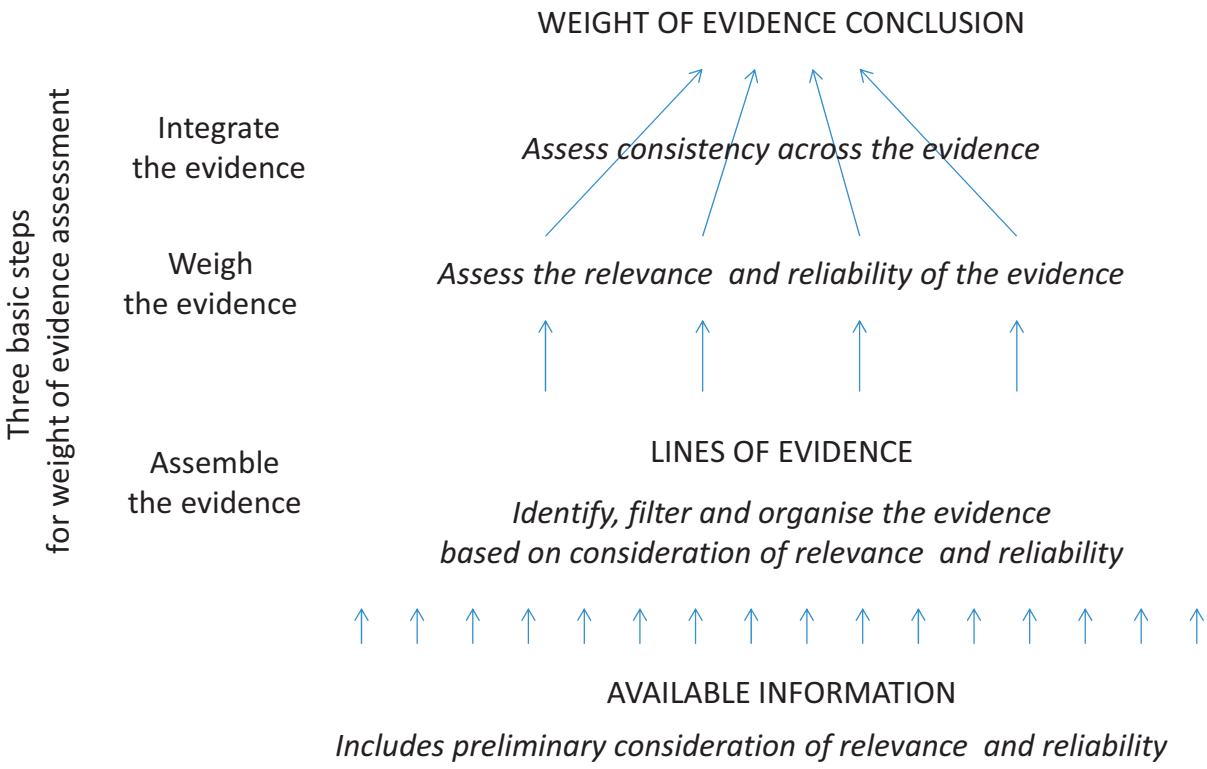

Figure 2: Relationship of relevance (including biological relevance), reliability and consistency to the three basic steps of weight of evidence assessment and to the conclusion for a weight of evidence question

The present document defines reliability, relevance and consistency, in terms of their contributions to the weight of evidence assessment, as illustrated in Figure 2:

Reliability is the extent to which the information comprising a piece or line of evidence is correct, i.e. how closely it represents the quantity, characteristic or event that it refers to. This includes both accuracy (degree of systematic error or bias) and precision (degree of random error).

Relevance is the contribution a piece or line of evidence would make to answer a specified question, if the information comprising the evidence was fully reliable. In other words, how close is the quantity, characteristic or event that the evidence represents to the quantity, characteristic or event that is required in the assessment. This includes biological relevance (EFSA, 2017) as well as relevance based on other considerations, e.g. temporal, spatial, chemical, etc.

Consistency is the extent to which the contributions of different pieces or lines of evidence to answering the specified question are compatible.

These definitions are compatible with those found in other publications relating to weight of evidence assessment (e.g. ECHA, 2010; SCENIHR, 2012; Vermeire et al., 2013). Different types of 'contribution' are discussed further below.

Other publications mention additional considerations relevant to weight of evidence assessment including, for example, quality, applicability, coherence, risk of bias, specificity, biological concordance or plausibility, biological gradient and many others. Some of these are synonyms for reliability, relevance or consistency, some refer to combinations of reliability and relevance, and others refer to specific types of reliability, relevance or consistency that are important for particular areas of assessment (see Annex B). For example, in some approaches for assessing reliability, emphasis is placed on particular aspects (e.g. conformance to GLP) with less consideration of other key aspects of reliability that should also be assessed, such as bias (Moermond et al., 2017). The reasons for giving particular emphasis to reliability, relevance and consistency are that they are generic considerations, applicable to every type of assessment, with defined relationships to the three basic steps of weight of evidence assessment, and a defined relationship to the weight of evidence conclusion, as illustrated in Figure 2. This makes them useful to assessors as a conceptual framework for identifying specific considerations relevant to particular assessments, and for assessing how they combine to determine the weight of evidence conclusion. How this can be applied in practice is discussed further in Section 4.

2.6. Relation of weight of evidence assessment and uncertainty

Weight of evidence assessment and uncertainty are closely related. For example, SCENIHR (2012) state that 'strength of evidence is inversely related to the degree of uncertainty', while Suter and Cormier (2011) state that 'the weight of the body of evidence, based on the combined weights of individual pieces of evidence, may be used to express confidence or uncertainty in the results'.

Weight of evidence assessment is defined above as a process which determines the relative support for possible answers to a scientific question. EFSA (2016a–c) defines uncertainty as 'a general term referring to all types of limitations in available knowledge that affect the range and probability of possible answers to an assessment question'. Answers may refer to alternative hypotheses or estimates, and probability is one way of expressing relative support for possible answers. Thus, the weight of evidence conclusion for a question and the uncertainty of the answer can be expressed in identical form: the range of possible answers and their relative degree of support or probability. This is explicit in meta-analysis, an evidence integration method producing conclusions in the form of estimates with confidence intervals, which express uncertainty.

However, expression of uncertainty for the conclusion of the scientific assessment as a whole should additionally include any uncertainties associated with the weight of evidence process itself. This may include uncertainties regarding, for example, the selection of evidence, assessment of reliability, relevance and consistency, and choice of weight of evidence methods. These should be taken into account by uncertainty analysis following the weight of evidence assessment (as indicated in Figure 1 above), and may modify the range and probability of answers to some degree.

Consistent with this, each of the three basic considerations in the weight of evidence assessment may be expressed in terms of uncertainty:

- the reliability of a piece or line of evidence can be expressed as the uncertainty of that evidence itself (i.e. how different the evidence might be if the information comprising it was correct);
- the relevance of a piece or line of evidence can be expressed as the uncertainty that would be associated with extrapolating from fully reliable information, of the type provided by that evidence, to its contribution to answering the weight of evidence question;
- limitations in consistency between pieces or lines of evidence add to uncertainty about the answer to the weight of evidence question. In principle, different pieces or lines of evidence for the same question should be consistent, if allowance is made for their reliability and relevance (this is discussed further in Section 4.4).

Thus weight of evidence assessment could be regarded as contributing to uncertainty analysis, addressing the part of uncertainty that relates to limitations in reliability, relevance and consistency of the evidence. Uncertainty which is addressed in the weight of evidence assessment does not need to be re-analysed in the uncertainty analysis, but may be added to (see Section 4.5).

Probability is not the only way to express relative support or uncertainty. It can also be expressed qualitatively, and this is essential for any uncertainties or aspects of weight of evidence assessment that cannot be quantified (see Section 5.10 of EFSA, 2016a–c). As explained above and by EFSA (2016a–c), expressing the probability of possible answers is important for the conclusions of a scientific assessment, but need not apply to earlier steps in the weight of evidence process.

2.7. Relation of weight of evidence assessment and variability

It is important also to consider how the weight of evidence assessment relates to variability. Variability is defined by EFSA (2016a–c) as 'heterogeneity of values over time, space or different members of a population, including stochastic variability and controllable variability'. Note that 'values' could refer to values on a quantitative scale, or to alternative qualitative descriptors, and that 'population' is not restricted to biological populations but may also refer to other entities (e.g. variability in temperature at different points in time and space).

Variability is often important in a scientific assessment, e.g. variability in chemical occurrence in food and consumption are important in chemical exposure assessment. This needs to be dealt with when defining the questions for assessment, such that they refer to specific descriptors or summaries of the variable quantity, such as the average or 95th percentile exposure. If weight of evidence assessment was used as part of the exposure assessment, for example to integrate occurrence data from different countries, the reliability, relevance and consistency of the different pieces or lines of

evidence regarding the variability of occurrence would be assessed. Thus, variability can be the *subject* of a weight of evidence assessment, rather than a contributor to it.

Variability in data is a combination of real variability of the quantity being measured (e.g. between individuals) and variability of the measurement process (measurement error, a form of uncertainty). Variability in data due to measurement error should be taken into account when assessing reliability. Variability between results reported by different studies might reflect differences in the reliability of those studies or differences in their relevance for the assessment. Such differences may lead to apparent inconsistencies in data that need to be considered when integrating evidence (Section 4.4).

3. Overview of qualitative and quantitative methods for weight of evidence assessment

3.1. Examples of weight of evidence approaches

3.1.1. Classification of weight of evidence approaches

Several reviews of weight of evidence approaches have been published, especially by Chapman et al. (2002), Weed (2005), Linkov et al. (2009), Lorenz et al. (2013) and ANSES (2016).

Instead of conducting another review of weight of evidence approaches, the SC scrutinised the existing reviews to identify approach(es), which could be useful for classifying weight of evidence methods suitable for ecotoxicological assessments. A classification proposed by Linkov et al. (2009) has been taken as a starting point (with modifications, see below) as it covers a broad range of methods with contrasting levels of complexity; it enables methods to be grouped according to whether they are qualitative and/or quantitative; and it can also capture weight of evidence methods identified in the more recent reviews. It is important to stress that the classification here is illustrative rather than prescriptive. Linkov et al. (2009) distinguish the following types of weight of evidence assessment:

Listing evidence: Presentation of individual lines of evidence without attempt at integration
Best professional judgement: Qualitative integration of multiple lines of evidence
Causal criteria: A criteria-based methodology for determining cause and effect relationships
Logic: Standardised evaluation of individual lines of evidence based on qualitative logic models
Scoring: Quantitative integration of multiple lines of evidence using simple weighting or ranking
Indexing: Integration of lines of evidence into a single measure based on empirical models
Quantification: Integrated assessment using formal decision analysis and statistical methods.

In the Linkov et al. (2009) classification system, approaches are grouped by the degree to which they are quantitative. The least quantitative approaches are categorised as 'Listing evidence', while the most quantitative ones fall within the category 'Quantification' and are based on statistical approaches or on formal decision-analytical tools. Other categories correspond to intermediate situations.

The current guidance does not formally distinguish between the categories 'listing evidence' and 'best professional judgement'. Linkov et al. (2009) suggest that approaches described as 'listing evidence' do not attempt to integrate lines of evidence together. Rather, lines of evidence 'are simply presented, although the assessor will at times make claims that the weight of evidence assessment points to specific conclusions' (Linkov et al., 2009). In 'listing evidence', it is implicit that integration of evidence must take place in order to reach conclusions, although the integration may not be based on a formal method. 'Listing evidence' is therefore considered here as being synonymous with 'best professional judgement' rather than being a separate category.

In addition, the current guidance groups together the Linkov et al. (2009) categories 'Logic', 'Scoring' and 'Indexing' as they are approaches involving rating which share many common characteristics. These three categories are considered here as a single category named 'Rating'. Thus, the SC considers four categories of weight of evidence assessment methods: best professional judgement, causal criteria, rating and quantification. Examples of these approaches are listed below for each category. Some of them cover all three basic steps of the general weight of evidence process, while others are more specific and focus on one or two steps. Approaches restricted to problem formulation were considered as outside the scope of this guidance, and are not included.

The SC does not aim to cover here all existing approaches, but rather to give a brief overview of different types of approaches and to provide key references. Examples are briefly presented in the following sections (note that these are included only to illustrate the categories and are not intended

as examples of best practice). Categories described in Sections 3.1.2–3.1.3 (best professional judgement and causal criteria) are collectively referred to in this guidance as qualitative approaches. Several criteria are then presented in Section 3.2 to help risk assessors to choose among weight of evidence approaches. These approaches have been used in different EFSA scientific assessments (see for example Annex C). Note that assessments may use a combination of more than one of these categories.

3.1.2. Category 'Best professional judgement'

In this category, no formal method is used for evidence integration. Instead, the listed pieces of evidence are used to form a conclusion by professional opinion via a discussion of the findings. The approaches of this category simply list pieces of evidence in text or in tables. The origin of the evidence depends on the approach used for assembling evidence, for example:

- Several methods of evidence synthesis which do not involve quantitative integration could be classified under this category. These include extensive literature searches (EFSA, 2010a; Higgins and Green, 2011), systematic maps (CEE, 2016; James et al., 2016), and non-quantitative systematic reviews (i.e. those lacking a quantitative data synthesis step) (EFSA, 2010a; Higgins and Green, 2011).
- Evidence can be directly provided by applicants in reports and/or datasets rather than being selected by risk assessors.

Clear documentation of the discussion is important to ensure transparency in how decisions were reached. Several examples are given in the review of Linkov et al. (2009), notably Staples et al. (2004).

3.1.3. Category 'Causal criteria'

Approaches of this category provide a structure based on explicit criteria to evaluate relationships between cause and effect from one or several lines of evidence. The Bradford Hill considerations (Hill, 1965) are widely used, especially in epidemiology. They are often seen as the minimal conditions needed to establish a causal relationship between two items, and are frequently used in epidemiological studies to assess the extent of supporting evidence on causality. They are also used, in modified form, to weight of evidence assessment in mode of action analysis (Meek et al., 2014). Several variants have been proposed in the literature. For example, Becker et al. (2015) proposed a template based on modified Bradford Hill considerations for weight of evidence assessment of adverse outcome pathways.

3.1.4. Category 'Rating'

This category includes a variety of frameworks that involve rating of evidence. Examples include GRADE[3] (Guyatt et al., 2011), WHO-IARC (IARC, 2006), WCRF/AICR (2007), OSHA (2016) and guidance used to produce NTP Monographs (NTP, 2015; OHAT, 2015). Guidance to assess and integrate evidence is based on several factors, often derived from the Bradford-Hill considerations. However, 'Rating' builds on the approaches of the category 'Causal criteria' in that more guidance is provided for appraising and integrating evidence.

These approaches usually relate to the second step of the weight of evidence process, including the appraisal of individual studies and rating confidence in the individual lines of evidence (e.g. 'high confidence,' 'sufficient evidence'). Some of them also provide tools based on a matrix for integrating lines of evidence to reach hazard identification conclusions (WHO-IARC, OHAT, OSHA (2016)). None of these approaches use formal probabilistic techniques, but it is possible to combine application of the structured framework guidance with a more quantitative presentation of conclusions.

Some of these approaches (e.g. GRADE) are designed to be flexible for use in a variety of disciplines and able to be applied under different time and resource constraints in situations corresponding to different levels of urgency (Thayer and Schünemann, 2016).

In EFSA, the scheme presented in the 'Guidance on a harmonised framework for pest risk assessment and the identification and evaluation of pest risk management options' (EFSA, 2010a) belongs to this category.

[3] GRADE stands for Grading of Recommendations, Assessment, Development and Evaluation.

3.1.5. Category 'Quantification'

This category covers a large diversity of approaches that can be used to integrate evidence into lines of evidence and/or to integrate different lines of evidence in order to reach a general conclusion.

This category includes standard statistical models such as fixed-effect and random-effect linear and generalised linear models. These are commonly used for meta-analysis. A typical application is to estimation of mean effect sizes, which can be interpreted as summary estimates of a quantity based on statistical integration of evidence from multiple primary studies. These statistical models may also be used for meta-regression to explain the variability between studies as a function of explanatory variables, for example, population characteristics or study quality issues. They are able to describe uncertainties through confidence intervals and probability distributions. Other types of statistical methods (e.g. Bayesian methods) are also useful for synthesising multiple sources of evidence.

In addition to statistical methods, other approaches have been proposed, especially machine learning, *in silico* tools and multicriteria analysis. Linkov et al. (2015) consider that multicriteria decision analysis can be used as a proxy for the Bayesian approach to weight of evidence assessment when model formulation is restricted by data limitations.

When a quantitative model is used for weight of evidence assessment, several authors recommend performing a sensitivity analysis to study the stability of the main conclusions to the model assumptions, e.g. to the model equations, or to the parameter values (Borenstein et al., 2009; Linkov et al., 2011).

Examples of quantitative approaches and several key references are listed below:

- Statistical methods for integrating data provided by several studies sharing similar characteristics (classic fixed-effect and random-effect models used in meta-analysis, and Bayesian hierarchical models). Many textbooks and methodological papers are available on these methods, for examples Borenstein et al. (2009) on classic techniques, and Sutton and Abrams (2001) and Higgins et al. (2015) on Bayesian methods in the context of meta-analysis.
- Statistical methods for integrating different types of studies in order to allow decisions based on all available evidence and to analyse uncertainty (Small, 2008; Turner et al., 2009; Gosling et al., 2013).
- Quantitative expert judgement including multicriteria decision analysis for integrating different types of studies (Linkov et al., 2011, 2015).
- Machine learning techniques (Li and Ngom, 2015).
- *In silico* tools including QSAR, PBTK-TD (ECHA, 2016).

3.2. Choosing weight of evidence methods

A challenge when planning a weight of evidence assessment is to determine which assessment method(s) to select, given the variety of different methods available. A single easy-to-use weight of evidence method that covers all the basic steps of the weight of evidence process and enables transparent quantification of uncertainty may not be available. A pragmatic approach is therefore recommended for identifying the most suitable method, or combination of methods, for the weight of evidence assessment. A list of criteria for comparing weight of evidence methods is suggested in Table 1, to assist in evaluating the relative strengths and weaknesses of the different methods. This list is not exhaustive but based on discussions during the development of the current guidance; the criteria have not been formally tested and are not intended to be prescriptive or mandatory, but are intended to aid the justification and selection of methods. The criteria may also be helpful for transparently recording and reporting the decision-making process used for weight of evidence method selection, in keeping with EFSA's requirement for transparency in the conduct and reporting of scientific assessments (EFSA, 2006).

As with any type of evidence synthesis, weight of evidence methods face a potential trade-off between what would be ideal in terms of resource requirement (i.e. rapid, cheap, methods) and scientific rigour (i.e. methods that transparently display uncertainty at all steps of the weight of evidence process). Careful consideration will be needed early on in the planning process (in problem formulation) to ensure that adequate resource (time, staff expertise) is available to achieve the desired level of scientific rigour. In EFSA weight of evidence assessments which have prespecified and fixed resources, the criteria in Table 1 could be used to judge the optimal scientific rigour that could be achieved within the available resources. Alternatively, if resource availability for a weight of evidence assessment is negotiable, or if a weight of evidence assessment is at a preliminary scoping phase, these criteria may be helpful for estimating the resource needs for the assessment.

Table 1: Criteria for assessing the relative strengths and weaknesses of weight of evidence methods

Criterion	Key considerations
Time needed	Weight of evidence methods that can be conducted quickly would be preferable where urgent weight of evidence assessments are required. However, rapid methods risk sacrificing scientific rigour. When estimating the time required for a particular weight of evidence method, consideration should be given to the availability of expertise, since this could influence the time required for a weight of evidence assessment
Amount and nature of the evidence	The amount of evidence and its type (e.g. experimental, expert knowledge, surveys, qualitative, quantitative or combination) may affect which methods of weight of evidence assessment could be applied: the choice among these would then be determined by other criteria in this table. The similarity of the available studies may also be relevant, e.g. in deciding whether meta-analysis is an option
Availability of guidance	Guidance on the weight of evidence method should be readily available in the public domain and, ideally, should be endorsed, e.g. through peer-review and/or wide acceptance. Guidance should document the rationale of the method, the full process, and how to interpret the results. Ideally, access to help and support facilities should be readily available (e.g. any relevant tools such as tutorials, software programs or modules). Availability of guidance is important both for the conduct and the critical appraisal of weight of evidence assessments. Weight of evidence methods that lack adequate guidance would rate poorly on this criterion
Expertise needed	Weight of evidence methods are likely to vary in the level of technical skill required to conduct them. Some quantitative methods, for example, may require specific skills in statistics and/or programming. The expertise requirement should be considered in relation to availability of expertise and tools and, if necessary, whether available resources would support the outsourcing of expertise or provision of training. The level of expertise required has implications both for the conduct and the critical appraisal of weight of evidence assessments
Transparency and reproducibility	Transparency and reproducibility are fundamental principles required by EFSA in its scientific assessments. Transparency should apply to all parts of the weight of evidence method, meaning that it should be possible to follow clearly how the input data for the assessment are analysed to produce the conclusions. Reproducibility is defined such that consistent results should be expected if the same method were to be repeated using the same input data (but note that results are unlikely to be identical, dependent on the degree to which expert opinion is involved)
Variability and uncertainty	Weight of evidence methods should, ideally, explicitly report and analyse both variability and uncertainty at all steps of the assessment, and propagate them appropriately through the assessment. Quantitative expression of variability and uncertainty is preferable to qualitative expression. Careful consideration may be needed to ensure that the weight of evidence method can include all relevant sources of variability and uncertainty
Ease of understanding for assessors and risk managers	Weight of evidence methods are likely to vary in how easy they are to understand by non-specialists, and this may be related to the expertise needed, as well as the availability of adequate guidance. It will be beneficial if the principles of the methods chosen can be readily understood by assessors and risk managers

The criteria in Table 1 should be considered together by assessors when planning weight of evidence assessment. This is because the strengths and weaknesses of weight of evidence methods are multidimensional, and individual criteria alone may not be able to capture important trade-offs, e.g. between resource availability and scientific rigour. Note that the criteria do not necessarily have equal importance: their relative importance may be discussed on a case-by-case basis when planning each weight of evidence assessment. The criteria in Table 1 are not exhaustive. Other criteria which may be useful include the strength and scope of the theoretical basis for a method and the extent to which the output of the method is in a form which can be tested.

4. Practical guidance for conducting weight of evidence assessment

This section contains practical guidance for applying weight of evidence approaches within EFSA scientific assessments. Assessors should choose the specific approaches that are best suited to the needs, time/resources available and context of their assessments.

Three types of assessment are distinguished, which require different approaches:

- assessments where the approach to integrating evidence is fully specified in a standardised assessment procedure;
- case-specific assessments, where there is no pre-specified procedure and assessors need to choose and apply approaches on a case-by-case basis;
- emergency procedures, where the choice of approach is constrained by unusually severe limitations on time and resources.

Standardised assessment procedures have been established in many areas of EFSA's work, especially for regulated products. Standardised procedures are generally defined in documents, e.g. EU regulations or EFSA guidance documents. They specify what questions should be addressed, what evidence is required, and what methods of assessment should be applied to it. They generally include standardised elements that are assumed to provide adequate cover for uncertainty (EFSA, 2016a–c).

Where a standard assessment involves questions that require integration of evidence, the methods for doing this will be specified in the standard procedure. For example, in human health risk assessment of chemicals in food, the outcome may often be based on one of the available studies, which is considered to provide the highest level of protection for the consumer. While not generally thought of as weight of evidence, this is a procedure for integration which, after considering all the evidence, in effect gives all the weight to a single study (sometimes referred to as the critical study).

In assessments following a standardised procedure, the default approach should be to integrate evidence using the methods as specified by the procedure. If the methods are specified in detail, they may be sufficient to conduct the assessment without further guidance; where the methods are not fully specified, the assessor may benefit from the guidance in Sections 4.1–4.6. If an assessment that would normally be addressed by a standard procedure includes weight of evidence issues that are not adequately addressed by the standard procedure, a case-specific approach will be needed for that part of the assessment, following the guidance in Sections 4.1–4.6.

In **case-specific assessments**, for which there is no standard procedure, evidence integration will need to be conducted case-by-case, following the guidance in Sections 4.1–4.6.

Emergency assessments are required in situations where there are exceptional limitations on time and resources. If an emergency assessment involves scientific questions that require integration of evidence, assessors should first consider whether any standard procedure exists that can be applied within the time and resources available. If not, then the assessor should conduct a case-specific assessment, choosing options from Sections 4.1–4.6 that are compatible with the time and resources available.

In some cases, the literature available for an assessment includes **previous reviews or assessments of a similar question**, which themselves involve weight of evidence assessment. Ideally, assessors should access and evaluate the original evidence used in the previous assessments, rather than treating the outcomes of previous assessments as evidence *per se*. However, when time and resources are too limited to access all the original evidence, then it may be justifiable to make use of the previous assessments in the new assessment. If this is done, it will be essential to take account of any differences between the questions addressed in the previous and new assessments, and of any differences or shortcomings in the criteria and assessment methodology that were used, and to document these considerations transparently.

4.1. Define the questions for weight of evidence assessment

A single assessment may comprise one or more scientific questions and none, some or all of those questions may require weight of evidence assessment. Interpreting the questions posed by the Terms of Reference, and deciding whether to subdivide them, is part of the first stage of scientific assessment, often referred to as problem formulation. General guidance on problem formulation for EFSA's scientific assessments is provided in other documents (EFSA, 2006, 2015a,b), and the need to ensure questions are well-defined is further discussed by EFSA (2016a–c). Weight of evidence assessment does not involve any additional requirements or considerations for specifying the questions for assessment, so the reader is referred to the documents referred to above for details.

Problem formulation also includes planning the strategy and methods for assessment (EFSA, 2015a, b). As part of this, the assessors should identify which of the questions in the assessment will require integration of evidence and therefore the use of weight of evidence approaches.

The output of problem formulation should therefore include a list of the questions. Each question that requires weight of evidence assessment should then be addressed by applying the basic steps

described in the three following subsections. When the assessment involves a hierarchy of questions, start with the questions at the lowest level of the hierarchy, as the conclusions of these will inform lines of evidence for higher questions. It is sometimes necessary to return to and revise the problem formulation later, if additional questions are identified in the course of the assessment.

4.2. Assemble the evidence

This is the first of the three basic steps of weight of evidence assessment for an individual question (see Section 2.4). Guidance for this and the following two steps is provided as a series of numbered (sub)steps, which may be considered in sequence. For every step, assessors should choose approaches that are appropriate to the needs and context of the assessment in hand, including any limitations on time and resources.

1) **Identify potentially relevant evidence**. In some EFSA assessments, the evidence to be used is defined by regulations or guidance, and/or submitted by applicants. This applies especially in standard assessment procedures. When data gaps (absence of required data) are identified, it may be possible to mitigate their effect using other evidence, for example by read-across, if this is permitted by the relevant regulations or guidance. In non-standard (case-specific) assessments, the assessors define the strategy and criteria for identifying and accessing potentially relevant evidence. Procedures for this, with varying degrees of formality, are described in EFSA guidance on extensive literature searching and systematic review (e.g. EFSA, 2010a, 2015b), which is designed to increase coverage and reduce potential biases in evidence gathering.
2) **Select evidence to include in the weight of evidence assessment**. In principle, all evidence identified as potentially relevant in step 1 should be taken into account, but limitations on time and resources may require the assessment to focus primarily on the most relevant and/or most reliable evidence. This subset of evidence may be identified by filtering or screening using appropriate criteria for relevance and/or reliability (EFSA, 2010b, 2015b). Evidence that was considered potentially relevant but not included should be retained separately, for example as a list of references or archive of documents, so that the impact of excluding it can be considered as part of uncertainty analysis (below).
3) **Group the evidence into lines of evidence**, i.e. subsets of evidence which the assessors find useful to distinguish when conducting the assessment. There are no fixed rules for how to form lines of evidence, but it may be helpful to distinguish those which are standalone and those that are complementary (Section 2.1). If the lines of evidence are complementary, they may be grouped according to the contribution they make to answer the question (e.g. exposure, hazard, etc.). Standalone lines of evidence may comprise evidence on the same aspect of the assessment but generated by different methods (e.g. different study types), with different subjects (e.g. species, chemicals, etc.) and in different conditions. This will tend to group evidence that has similar relevance and/or reliability. The lines of evidence and the rationale for constructing them should be documented, identifying which are standalone and which are complementary.

4.3. Weigh the evidence

Four broad categories of methods for weight of evidence assessment are presented in Section 3, together with suggestions for choosing between them: best professional judgement, causal criteria, rating and quantitative methods. Assessors should first consider the possibility of using quantitative methods because an appropriate and well-conducted quantitative analysis will generally be more rigorous than other methods. For example, when it is possible and appropriate to combine multiple studies by meta-analysis, this will be more rigorous than integrating them by expert judgement. However, there are two important caveats to this. First, quantitative methods may not be appropriate for various reasons, e.g. not applicable to the nature, quantity or heterogeneity of the evidence to be integrated, not practical within the time and resources available, etc. Second, quantitative methods may not address all the considerations that are relevant for weighing the evidence. For example, common approaches to meta-analysis only capture those aspects of reliability and consistency that are represented in the variability of the data, although some forms of meta-analysis can also take account of relevance and additional aspects of reliability (e.g. Turner et al., 2009).

Therefore, the approach proposed below is to check first whether quantitative methods are practical and appropriate, and then complement them with qualitative methods (categories to ensure all relevant considerations are addressed). When quantitative methods are not practical or appropriate, only qualitative methods can be used. However, assessors should start by deciding what considerations are relevant for weighing the evidence, as this may have implications for the choice of methods.

When there are data gaps, due to the absence of data that are normally required, the weight that those data would have had will be absent and this will be taken into account when the available evidence is integrated (see Section 4.3).

1) **Decide what considerations are relevant for weighing the evidence**. The general considerations for weighing evidence are reliability, relevance and consistency, as explained and defined in Section 2.5. Assessors may choose to work with these three basic considerations, or use more specific criteria appropriate to their area of work, especially if these have already been established in guidance or the scientific literature. If using pre-established criteria, assessors should check that they cover all aspects of reliability, relevance and consistency that are relevant for the assessment in hand, and define any additional criteria that are needed.
    a) **Reliability** is the extent to which the information comprising a piece or line of evidence is correct. It may be assessed by considering the uncertainty of the evidence, i.e. how different it might be if the information comprising it was correct. Everything that contributes to that uncertainty should be included when assessing reliability.
    b) **Relevance** is the contribution a piece or line of evidence would make to answer a specified weight of evidence question, if the information comprising the evidence were fully reliable. Everything that contributes to the need for extrapolation, and its uncertainty, should be included when assessing relevance.
        i) For a standalone line of evidence, consideration of relevance involves thinking about how well that evidence would answer the question, if the information comprising it were fully reliable. How much extrapolation is involved, between the subjects and conditions the evidence relates to and those relevant for the question and how uncertain is that?
        ii) For a complementary line of evidence, consideration of relevance involves identifying what that evidence contributes to the conceptual model or argument for answering the question, and considering what extrapolation is required to provide that contribution.
    c) **Consistency** should be considered when integrating evidence (below).
2) **Decide on the method(s) to be used for weighing and integrating the evidence**. Refer to the categories and criteria in Section 3. Some of the methods for weighing evidence also perform the integration step (e.g. meta-analysis), or limit the choice of methods for integration, so both steps should be considered when choosing between methods. The choice of methods may also be affected by whether the lines of evidence are standalone or complementary. For example, meta-analysis can be used to integrate standalone lines of evidence, whereas complementary lines of evidence require a quantitative model of the relationships between the lines of evidence and the answer to the question (e.g. the relationships between exposure, hazard and risk).
    a) **Consider whether quantitative methods are practical and appropriate for the needs and context of the assessment**. If it is decided to use a quantitative method, identify which aspects of reliability, relevance and consistency it will address, and which it will not.
    b) **Choose one or more qualitative methods to address those aspects of reliability, relevance and consistency that are not treated quantitatively**. This could include methods from one or more of the non-quantitative categories presented in Section 3.
    c) **Check that the chosen methods (quantitative and/or qualitative) address all pertinent aspects of reliability, relevance and consistency (identified in step 1 above)**.

d) **If more than one method is chosen for weighing evidence, consider whether their results can be combined directly when integrating the evidence**. Some methods are capable of incorporating the outputs of other methods: e.g. Doi (2014) has developed methods for incorporating quality scores into meta-analysis, while Turner et al. (2009) have proposed methods for incorporating quantitative expert judgements about the effects of study limitations (which may include reliability and relevance) into meta-analysis. Such methods may be used, if they are appropriate and practical for the needs and context of the assessment.

3) **Apply the chosen methods for weighing the evidence and summarise the results in a form that is helpful for integration**. Weighing will often be conducted at the level of individual pieces of evidence. Alternatively, pieces of evidence within the same line of evidence could be weighed collectively. The latter option may be quicker when time is limited but requires an implicit integration of the pieces within the line of evidence, so is less transparent and may be more challenging for the assessors to perform (because it requires weighing and integrating simultaneously).
When more than one method of weighing is used (e.g. a quantitative method combined with a qualitative method), it is recommended to find a way of presenting the results together in a concise tabular or graphical summary. For example, estimates and confidence intervals from quantitative methods can be plotted on a graph alongside symbols or text showing the results of qualitative methods (e.g. EFSA, 2015a, 2016a). This provides a useful overview of the evidence, which may be helpful for the assessors in the integration step and also for others, who read the finished assessment.

4.4. Integrate the evidence

In this step, the evidence is integrated to arrive at the conclusion, taking account of the reliability and relevance of the evidence, assessed in the preceding step, and also the consistency of the evidence. To reach a conclusion on the weight of evidence question, integration is necessary both within and between lines of evidence. When there are data gaps, due to the absence of data that are normally required, the absence of the weight those data would have contributed will be reflected in the outcome of the integration process. When appropriate, the effect of this may be mitigated by the contributions of other evidence (e.g. read-across), or taken into account by use of assessment factors (which should themselves be evidence-based).

1) **Consider the conceptual model for integrating the evidence.** Integration always involves a conceptual model, even if this is not made explicit. Integrating standalone lines of evidence requires a conceptual model of how evidence of differing weight is combined. Integrating complementary lines of evidence additionally requires a conceptual model of the contributions made by the different lines of evidence and how they combine to answer the question. In both cases, it is important to take account of any dependencies between different pieces and/or lines of evidence. Dependencies can have an important impact on how evidence should be integrated (see point 3 below).
Assessors may find it helpful to make the conceptual model explicit, e.g. as a flow chart or list of logical steps. This should help assessors to take appropriate account of the relationships and dependencies between pieces and lines of evidence, and between the evidence and the question being assessed, both when the integration of evidence is done by expert judgement and when it is done using a quantitative model. Making the conceptual model explicit also contributes importantly to the transparency of the assessment.
2) **Assess the consistency of the evidence**. Consistency is the extent to which the contributions of different pieces or lines of evidence are compatible (Section 2.5). Limitations in consistency arise in part from limitations in the relevance and reliability of different pieces or lines of evidence. If a question is well-defined, only a single correct answer should be possible, and any apparent inconsistencies in the evidence should be explicable in terms of differences in reliability and/or relevance. Assessors should not, however, simply conclude that inconsistent evidence is unreliable or irrelevant. Rather, assessors should consider whether, after taking differences in reliability and relevance into account, the pieces or lines of evidence still appear inconsistent. If so, this may imply the presence of additional limitations in relevance and reliability, beyond those already taken into account, or limitations in the conceptual model for

integrating the evidence. Alternatively, it may imply there is more than one possible answer to the question, in which case the question may need to be more precisely defined or split into two or more separate questions. Any remaining inconsistency should be considered as part of the uncertainty affecting the weight of evidence conclusion.

3) **Apply the method(s) chosen for integrating the evidence** (the methods chosen in step 2 of Section 4.3, above). As already mentioned, one or more methods may be used, from one or more of the categories of methods described in Section 3.
It may often be helpful to approach integration hierarchically, starting with evidence that is more closely related. This is one of the reasons for grouping evidence into lines: first, pieces of evidence are integrated within each line of evidence, and then different lines of evidence are integrated. When integrating evidence, it is important to take account of dependencies. The result of integrating two dependent pieces or lines of evidence will have more or less weight than had they been judged independent (see EFSA, 2016a–c for further discussion of dependency).
4) **Develop the conclusion for the weight of evidence assessment**. If there is no single method that can take into account all the pertinent aspects of reliability, relevance and consistency, it will be necessary to integrate the results of the different methods by expert judgement. This may be done within the process for reaching a conclusion, as follows:

   a) **Summarise all the results up to this point in a concise tabular or graphical format**. This should comprise all the results of weighing the evidence, as in step 3 of Section 4.3, together with the results of any integration that has been done (e.g. integrated estimates and confidence intervals from meta-analysis or integrated scores from scoring methods).
   b) **Define how the conclusion of the weight of evidence assessment (range of possible answers and their relative support) will be expressed**. The range of answers may be expressed on an appropriate quantitative scale or as alternative qualitative statements or propositions. Relative support for the possible answers may be expressed quantitatively, e.g. as probabilities or qualitatively.

      i) When weight of evidence assessment directly addresses the conclusion of a scientific assessment, the range of possible answers and how probable they are should be expressed quantitatively if possible. Any considerations that cannot be included in the quantitative expression imply that the conclusion will be subject to unquantified uncertainties, which should be described qualitatively (see EFSA, 2016a–c). If no quantitative expression is possible, this implies that each probability could be anywhere between 0% and 100%, and the assessor should consider whether the evidence supports any conclusion at all (see also Sections 5.10 and 5.11 of EFSA, 2016a–c).
      ii) When weight of evidence assessment addresses an intermediate question in a larger assessment, the range of answers and their relative support may expressed either qualitatively or quantitatively, depending what is convenient for use in subsequent steps of the assessment. One method could be taken as the primary method of integration, for use in subsequent steps of the assessment, and additional considerations (not covered by the primary method) could be carried over to the uncertainty analysis at the end of the scientific assessment as a whole. For example, if meta-analysis was used to integrate occurrence data in a chemical risk assessment, the output of the meta-analysis could be used in subsequent steps of exposure and risk assessment. Any other considerations regarding the quality and relevance of the occurrence data and the assumptions of the meta-analysis would be addressed as part of combined uncertainty at the end of the risk assessment (EFSA, 2016a–c, Section 12).

   c) **Where expert judgement is required, use an appropriate procedure for this**. Expert judgement should always be careful, reasoned, evidence-based and transparently documented. This may be achieved through formal expert knowledge elicitation (EFSA, 2014a), or semiformal expert knowledge elicitation or expert discussion (EFSA, 2016a–c). The cited documents focus on eliciting distributions for quantitative parameters, but the underlying principles can be applied also to eliciting probabilities for alternative answers

to a qualitative question. Assessors should choose a procedure that is appropriate for the needs, timeframe and context of their assessment. For example, if the judgement is likely to be critical for decision-making, that would be a reason for more formal methodology, if time and resources allow.

4.5. Uncertainty and influence analysis

All EFSA scientific assessments must include consideration of uncertainties, reporting clearly and unambiguously what sources of uncertainty have been identified and what their impact on the assessment outcome is. It is recommended that the combined impact of as many as possible of the identified uncertainties be expressed quantitatively, in terms of the range and probability of possible answers to the assessment question, and that any uncertainties that cannot be included in this should be described qualitatively (EFSA, 2016a–c). These recommendations apply to weight of evidence approaches, as well as other types of scientific assessment.

Weight of evidence assessment contributes to uncertainty analysis, as explained in Section 2.6. However, weight of evidence conclusions expressed as the range of possible answers and their relative support (e.g. an estimate and confidence interval from a meta-analysis) may not incorporate all the uncertainty affecting the weight of evidence assessment. First, the assessors may have omitted some considerations regarding the evidence from the integration process, leaving them to be addressed in the uncertainty analysis, as described in step 4b.ii. of Section 4.4. Second, there will often be additional uncertainties associated with the identification of evidence (including the choice of search criteria), the impact of any potentially relevant evidence that was excluded from detailed assessment, and the choice and implementation of methods for assembling, weighing and integrating data. This includes 'uncertainties in the judgement used' in weight of evidence assessment (SCENIHR, 2012). Assessors should systematically document all identifiable uncertainties affecting the weight of evidence assessment, and take them into account in the assessment of combined uncertainty for the overall scientific assessment (EFSA, 2016a–c, Section 12).

Influence analysis or sensitivity analysis is an optional part of scientific assessment (EFSA, 2016a–c). It can be valuable in identifying which sources of uncertainty contribute most to the uncertainty of assessment conclusions, and hence in targeting refinement of the assessment when this is required. When applied to assessments that include weight of evidence approaches, this could help decide whether and where to refine the weight of evidence assessment (see below). In meta-analysis, it is good practice to study the influence of individual primary studies on the effect size estimates (e.g. leverage plots, jack-knife procedure) and to study the impact of the modelling approach (e.g. random effect model vs. fixed effect model). In addition, meta-regression has been recommended to study the sensitivity of the effect size estimates to explanatory factors related to the study characteristics. In case of heterogeneity in the results of primary studies this may support identification of major sources of uncertainties and variability. Sensitivity analysis was also applied with other weight of evidence approaches, especially with quantitative multicriteria analysis. For example, the last step of the quantitative multicriteria analysis framework described by Linkov et al. (2011) explicitly deals with sensitivity analysis. It is useful to explore analogies between these formal approaches for influence analysis and similar approaches applicable in less formal methods for evidence integration. For example, it can be recommended to study the effect of leaving out individual lines of evidence or, if applicable, individual pieces (e.g. primary studies) on the weight of evidence conclusions. Likewise, alternative methods for evidence integrating could be used and their influence on the weight of evidence conclusion be demonstrated.

4.6. Iterative refinement of the assessment

Iterative refinement is an option in any type of scientific assessment. It is generally aimed either at reducing uncertainty or improving the characterisation of uncertainty, in those areas of the assessment that contribute most to the uncertainty of the assessment conclusions as identified by influence analysis or sensitivity analysis (EFSA, 2016a–c). In general, assessment should start at a level of refinement the assessors consider appropriate to the needs and context of the assessment, and then be refined as far as is necessary to inform decision-making or until the agreed time and resources are expended.

When refinement is needed in parts of an assessment where weight of evidence approaches are used, this can be achieved by returning to earlier steps of the process (illustrated in Figure 1), depending on what contributes best to refining the assessment. In some cases, it may be sufficient to

refine one or more of the basic steps of the weight of evidence assessment, whereas in other cases it may be beneficial to return to problem formulation and reformulate the questions to be addressed. For example, if a question involving complementary lines of evidence contributed significantly to overall uncertainty, consideration could be given to further subdividing the question, addressing each complementary part as a separate subquestion, and then combining their conclusions (with the option of using a quantitative model).

5. Reporting weight of evidence assessment

If the weight of evidence assessment has been conducted following a standardised procedure previously established for use in this area of EFSA's work, the weight of evidence assessment may be reported in the manner that is normal for that standardised procedure, provided this is transparent. The standardised procedure should be referenced and its applicability to the case in hand should be explained if it is not self-evident.

All other weight of evidence assessments should be reported as described below, although the level of detail may be reduced due to time constraints in emergency assessments.

Reporting should be consistent with EFSA's general principles regarding transparency (EFSA, 2006, 2009) and reporting (EFSA, 2014a, 2015b). In a weight of evidence assessment, this should include justifying the choice of methods used, documenting all steps of the procedure in sufficient detail for them to be repeated, and making clear where and how expert judgement has been used. Where the assessment used methods that are already described in other documents, it is sufficient to refer to those. Reporting should also include referencing and, if appropriate, listing or summarising all evidence considered, identifying any evidence that was excluded; detailed reporting of the conclusions; and sufficient information on intermediate results for readers to understand how the conclusions were reached.

Weight of evidence assessment is part of the wider process of scientific assessment, as illustrated earlier in Figure 1. Guidance on reporting other parts of the wider procedure, including evidence review, problem formulation and uncertainty analysis, is provided elsewhere (e.g. EFSA, 2014b, 2015b, 2016a–c). This section focusses on the reporting for the three basic steps of weight of evidence assessment: assembling the evidence, weighing the evidence and integrating the evidence. These steps should be reported separately for each scientific question or subquestion that is addressed.

To aid transparency and accessibility for readers it may be useful to summarise weight of evidence assessment in a tabular form. A suggested format is shown in Table 2. If a tabular format is not used, then all the information listed in Table 2 must be included in the assessment report, in a location and format that can easily be located by the reader (e.g. identifiable from section headings in the table of contents). If the information is presented in tabular form, it should be concise (ideally not more than 1 page per table) and refer the reader to the text of the opinion for details.

Table 2: Optional tabular format for summarising weight of evidence assessment

Question		***Insert text of question here***
Assemble the evidence	Select evidence	*Briefly summarise the methods used to search, select and extract the evidence (see Note 1)*
	Lines of evidence	*List the line(s) of evidence into which the evidence were assembled for assessment and identify any that are missing (see Note 2)*
Weigh the evidence	Methods	*Briefly summarise the method(s) used to weigh the pieces and lines of evidence (see Note 3)*
	Results	*Give a reference to the section of the assessment where the results of weighing the pieces and lines of evidence are presented (see Note 4)*
Integrate the evidence	Methods	*Briefly summarise the methods used to integrate the pieces and lines of evidence (see Note 5)*
	Results	*State the conclusions of integrating the evidence for this question (see Note 6)*

Italic descriptions are for guidance only and should be deleted once the table is completed.
Notes cited in the table are presented below.

Notes to Table 2:
Note 1. The summary of the methods used to search, select and extract the evidence should include, for example whether an extensive literature search or systematic review was conducted, and whether any of the evidence was obtained by expert elicitation and if so by which method.
Note 2. When listing the lines of evidence, give enough information for the reader to understand what they contain and how they differ. Present them as numbered bullets for ease of reference. State whether the lines of evidence are complementary, or standalone, or a mixture of both (see Section 2.1). Identify lines of evidence that were generated by (are conclusions from) preceding weight of evidence questions (if any). Also, identify any lines of evidence that are required (e.g. by legislation or guidance documents) but missing, i.e. data gaps.
Note 3. When summarising the method(s) used to weigh the pieces and lines of evidence, give enough information to make clear the type of method involved (see types of method in Section 3). If weighing and integration was done in preceding subquestions, refer the reader to where that is described. Refer the reader to the sections of the assessment where details of each method are provided.
Note 4. The detailed results of weighing the evidence must be presented together, in an appropriate part of the assessment report, in a format that helps the reader to compare the results for the different pieces and lines of evidence (e.g. a tabular listing). If they can be summarised briefly, include them in Table 2.
Note 5. Briefly summarise the methods used to integrate the pieces and lines of evidence, giving enough information to make clear the type of method involved (see types of method in Section 3 of Guidance).
Note 6. State the conclusion of integrating the evidence for this question in a form that expresses the range of possible answers and their relative support.

- If the weight of evidence assessment directly addresses the conclusion of a scientific assessment, results of weighing and integration that have been conducted by different methods (e.g. a combination of qualitative and quantitative methods), should integrated into a single conclusion on the relative support for different answers to the question, and expressed quantitatively to the extent that is possible. Any considerations that remain unquantified should be described qualitatively.
- If the weight of evidence assessment addresses an intermediate question in a larger scientific assessment, results of weighing and integration that were conducted by different methods may be expressed either qualitatively or quantitatively. They may either integrated into a single conclusion here, or carried forward separately to later stages of the scientific assessment.

6. Way forward and recommendations

This guidance document is intended to guide EFSA Panels and staff on the use of the weight of evidence approach in scientific assessments. It provides a flexible framework that must be used in all areas within EFSA's remit, within which assessors should apply those methods which most appropriately fit the purpose of their individual assessment.

This guidance is intentionally and necessarily a general framework. However, the SC believes that the principles and process are clear enough for EFSA Panels to apply them in their respective areas of work. Where appropriate, this could lead to relevant approaches being incorporated in area-specific guidance documents.

EFSA Panels and Units must ensure that their existing approaches to weight of evidence assessment are in line with the unconditional requirements of the current guidance document (Section 1.5). In particular, they should consider:

- whether all aspects of reliability, relevance and consistency that are pertinent in their area of work are addressed,
- how to ensure the transparency of weight of evidence assessments.

It is further recommended that:

- EFSA identify areas of its work where a formalised weight of evidence assessment is especially needed and initiate further work to apply suitable approaches from this guidance in those areas. This might include, for example, the integration of different types of evidence in chemical risk assessment, including *in vivo*, *in vitro*, *in silico*, omics, PBPK modelling as well as the Mode of Action and Adverse Outcome Pathway concepts.
- EFSA identify specific weight of evidence approaches that may provide added value in EFSA's work (especially quantitative methods, e.g. meta-analysis) and consider whether further guidance or training on them would facilitate uptake.
- EFSA should explore how to apply weight of evidence approaches in rapid scientific assessments, where time and resources are limited.

In implementing all the aforementioned recommendations, it is suggested that EFSA continues to collaborate at the European and international level to harmonise developments in this area.

References

Ågerstrand M and Beronius A, 2016. Weight of evidence evaluation and systematic review in EU chemical risk assessment: Foundation is laid but guidance is needed. Environment International, 92–93, 590–596. https://doi.org/10.1016/j.envint.2015.10.008

ANSES, 2016. Evaluation du poids des preuves à l'Anses: revue critique de la littérature et recommandations à l'étape d'identification des dangers. Rapport d'expertise collective. Saisine 2015-SA-0089, 116 pp.

Becker RA, Ankley GT, Edwards SW, Kennedy SW, Linkov L, Meek B, Sachana M, Segner H, Van Der Burg B, Villeneuve DL, Watanabe H and Barton-Maclaren TS, 2015. Increasing Scientific Confidence in adverse outcome pathways: application of tailored bradford-hill considerations for evaluating weight of evidence. Regulatory Toxicology and Pharmacology, 72, 514–537.

Beronius A, Molander L, Rudén C and Hanberg A, 2014. Facilitating the use of non-standard in vivo studies in health risk assessment of chemicals: a proposal to improve evaluation criteria and reporting. Journal of Applied Toxicology, 34, 607–617. https://doi.org/10.1002/jat.2991

Borenstein M, Hedges LV, Higgins JPT and Rothstein HR, 2009. Introduction to meta-analysis. Wiley, Chichester, UK.

CEE (Collaboration for Environmental Evidence), 2016. Systematic mapping. Available online: http://environmentalevidence.org/wp-content/uploads/2014/05/EE_InstructionsforAuthors_SYSTMAPS.pdf

Chapman PM, McDonald BG and Lawrence GS, 2002. Weight-of-evidence issues and frameworks for sediment quality (and other) assessments. Human and Ecological Risk Assessment: An International Journal, 8, 1489–1515. https://doi.org/10.1080/20028091057457

Collier ZA, Gust KA, Gonzalez-Morales B, Gong P, Wilbanks MS, Linkov I and Perkins EJ, 2016. A weight of evidence assessment approach for adverse outcome pathways. Regulatory Toxicology and Pharmacology, 75, 46–57.

Doi SAR, 2014. Evidence synthesis for medical decision making and the appropriate use of quality scores. Clinical Medical Research, 12, 40–46. https://doi.org/10.3121/cmr.2013.1188

ECHA (European Chemicals Agency), 2010. Practical guide 2: how to report weight of evidence. ECHA, Helsinki, pp. 1–26.

ECHA (European Chemicals Agency), 2015a. Guidance on the Application of the CLP Criteria Guidance to Regulation (EC) No 1272/2008 on classification, labelling and packaging (CLP) of substances and mixtures Version 5.0 July 2017

ECHA (European Chemicals Agency), 2015b. ECHA Guidance on biocides legislation. Available from: https://echa.europa.eu/guidance-documents/guidance-on-biocides-legislation

ECHA (European Chemicals Agency), 2016. Practical guide. How to use and report (Q)SARs. Version. 3.1. Available online: https://echa.europa.eu/documents/10162/13655/pg_report_qsars_en.pdf

EFSA (European Food Safety Authority), 2006. Transparency in risk assessment carried out by EFSA: guidance document on procedural aspects. EFSA Journal 2006;4(5):353, 16 pp. https://doi.org/10.2903/j.efsa.2006.353

EFSA (European Food Safety Authority), 2009. Guidance of the Scientific Committee on transparency in the scientific aspects of risk assessments carried out by EFSA. Part 2: general principles. EFSA Journal 2009;7(5):1051, 22 pp. https://doi.org/10.2903/j.efsa.2009.1051

EFSA (European Food Safety Authority), 2010a. Application of systematic review methodology to food and feed safety assessments to support decision making. EFSA Journal 2010;8(6):1637, 90 pp. https://doi.org/10.2903/j.efsa.2010.1637

EFSA (European Food Safety Authority), 2010b. Guidance on a harmonised framework for pest risk assessment and the identification and evaluation of pest risk management options. EFSA Journal 2010;8(2):1495, 66 pp. https://doi.org/10.2903/j.efsa.2010.1495

EFSA (European Food Safety Authority), 2011. EFSA Scientific Committee; Statistical Significance and Biological Relevance. EFSA Journal 2011;9(9):2372, 17 pp. https://doi.org/10.2903/j.efsa.2011.2372

EFSA (European Food Safety Authority), 2012. Guidance on selected default values to be used by the EFSA Scientific Committee, Scientific Panels and Units in the absence of actual measured data. EFSA Journal 2012;10(3):2579 pp. https://doi.org/10.2903/j.efsa.2012.2579

EFSA (European Food Safety Authority), 2014a. Discussion Paper - Transformation to an "Open EFSA". Available online: www.efsa.europa.eu

EFSA (European Food Safety Authority), 2014b. Guidance on expert knowledge elicitation in food and feed safety risk assessment. EFSA Journal 2014;12(6):3734, 35 pp. https://doi.org/10.2903/j.efsa.2014.3734

EFSA (European Food Safety Authority), 2014c. Guidance on statistical reporting. EFSA Journal 2014;12(12):3908, 18 pp. https://doi.org/10.2903/j.efsa.2014.3908

EFSA (European Food Safety Authority), 2014d. Systematic review guidance.

EFSA (European Food Safety Authority), 2015a. Editorial: increasing robustness, transparency and openness of scientific assessments. EFSA Journal 2015;13(3):e13031, 3 pp. https://doi.org/10.2903/j.efsa.2015.e13031

EFSA (European Food Safety Authority), 2015b. Scientific report on principles and process for dealing with data and evidence in scientific assessments. EFSA Journal 2015;13(5):4121, 35 pp. https://doi.org/10.2903/j.efsa.2015.4121

EFSA (European Food Safety Authority), 2016a. Guidance on Uncertainty in EFSA Scientific Assessment - Draft version for internal testing. Available online: https://www.efsa.europa.eu/en/topics/topic/unvertainty

EFSA (European Food Safety Authority), 2016b. Scientific opinion on the evaluation of the safety and efficacy of Listex™ P100 for reduction of pathogens on different ready-to-eat (RTE) food products. EFSA Journal 2016;14(8):4565, 94 pp. https://doi.org/10.2903/j.efsa.2016.4565

EFSA (European Food Safety Authority), 2016c. Guidance to develop specific protection goals options for environmental risk assessment at EFSA, in relation to biodiversity and ecosystem services. EFSA Journal 2016;14(6):4499, 50 pp. https://doi.org/10.2903/j.efsa.2016.4499

EFSA ANS Panel (EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS)), 2012. Guidance for submission for food additive evaluations. EFSA Journal 2012;10(7):2760, 60 pp. https://doi.org/10.2903/j.efsa.2012.2760. Available online: www.efsa.europa.eu/efsajournal

EFSA Scientific Committee, 2013. Scientific Opinion on Priority topics for the development of risk assessment guidance by EFSA's Scientific Committee. EFSA Journal 2013;11(8):3345, 20 pp. https://doi.org/10.2903/j.efsa.2013.3345

EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen HK, More S, Younes M, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Schlatter JR, Silano V, Solecki R, Turck D, Bresson J-L, Griffin J, Hougaard Benekou S, vanLoveren H, Luttik R, Messean A, Penninks A, Ru G, Stegeman JA, dervan Werf W, Westendorf J, Woutersen RA, Barizzone F, Bottex B, Lanzoni A, Georgiadis N and Alexander J, 2017. Guidance on biological relevance. EFSA Journal 2017;15(8):4970, 73 pp. https://doi.org/10.2903/j.efsa.2017.4970

EPA, 2003. EPA 100/B-03/001 June 2003 U.S. Environmental Protection Agency A Summary of General Assessment Factors for Evaluating the Quality of Scientific and Technical Information Prepared for the U.S. Environmental Protection Agency by members of the Assessment Factors Workgroup, a group of the EPA's Science Policy Council.

Gosling JP, Andy Hart H, Owen David M, Li J and MacKay C, 2013. A Bayes linear approach to 25 weight-of-evidence risk assessment for skin allergy. Bayesian Analysis, 8, 169–186.

Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, DeBeer H, Jaeschke R, Rind D, Meerpohl J, Dahm P, Schunemann HJ, 2011. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology, 64, 383–394.

Higgins J and Green S, 2011. Cochrane 5 Handbook for Systematic Reviews of Interventions. Available online: http://handbook.cochrane.org

Higgins JP, Thompson SG and Spiegelhalter DJ, 2015. A re-evaluation of random-effects meta-analysis. The Journal of the Royal Statistical Society, 172, 137–159.

Hill AB, 1965. The environment and the disease: association or causation?. Proceedings of the Royal Society of Medicine Royal Society of Medicine, 58, 295–300.

Hull RN and Swanson S, 2006. Sequential analysis of lines of evidence--an advanced weight-of-evidence approach for ecological risk assessment. Integrated Environmental Assessment and Management, 2, 302–311.

IARC, 2006. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans - Preamble. World Health Organization International Agency for Research on Cancer, Lyon, France.

IPCS (International Programme on Chemical Safety), 2004. *IPCS Risk Assessment Terminology*. International Programme on Chemical Safety, WHO, Geneva.

James KL, Randall NP and Haddaway NR, 2016. A methodology for systematic mapping in environmental sciences. Environmental Evidence, 5, 1.

Li Y and Ngom A, 2015. "Data integration in machine learning," Bioinformatics and Biomedicine (BIBM). 2015 IEEE International Conference on, Washington, DC, 2015, pp. 1665–1671. https://doi.org/10.1109/bibm.2015.7359925. Available online: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7359925&isnumber=7359638

Linkov I, Loney D, Cormier S, Satterstrom FK and Bridges T, 2009. Weight-of-evidence evaluation in environmental assessment: review of qualitative and quantitative approaches. Science of the Total Environment, 407, 5199–5205.

Linkov I, Welle P, Loney D, Tkachuk A, Canis L, Kim JB and Bridges T, 2011. Use of multicriteria decision analysis to support weight of evidence evaluation. Risk Analysis, 31, 1211–1225.

Linkov I, Massey O, Keisler J, Rusyn I and Hartung T, 2015. From "weight of evidence" to quantitative data integration using multicriteria decision analysis and Bayesian methods. Altex, 32, 3–8.

Lorenz RR, Goodman JE, Bailey LA, Prueitt RL, Beck NB, Bevan C, Honeycutt M, Kaminski NE, Paoli G, Pottenger LH, Scherer RH, Wise KC and Becker RA, 2013. A survey of frameworks for best practices in weight-of-evidence analyses. Critical Reviews in Toxicology, 43, 753–784.

Meek ME (Bette), Palermo CM, Bachman AN, North CM and Lewis RJ, 2014. Mode of action human relevance (species concordance) framework: evolution of the Bradford Hill considerations and comparative analysis of weight of evidence. Journal of Applied Toxicology, 34, 595–606. Published online 2014 Feb 10. https://doi.org/10.1002/jat.2984

Moermond C, Beasley A, Breton R, Junghans M, Laskowski R, Solomon K and Zahner H, 2017. Assessing the reliability of ecotoxicological studies: An overview of current needs and approaches. Integrated Environmental Assessment and Management, 13, 640–651. https://doi.org/10.1002/ieam.1870

Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, Guyatt G, Hooijmans C, Langendam M, Mandrioli D, Mustafa RA, Rehfuess EA, Rooney AA, Shea B, Silbergeld EK, Sutton P, Wolfe MS, Woodruff TJ, Verbeek JH, Holloway AC, Santesso N and Schünemann HJ, 2016. GRADE: Assessing the quality of evidence in environmental and occupational health. Environment International, 92-93, 611–616. https://doi.org/10.1016/j.envint.2016.01.004. Epub 2016 Jan 27.

National Research Council (U.S.) Committee on Improving Risk Analysis Approaches Used by the U.S. EPA, 2009. Science and decisions: Advancing risk assessment/ Committee on Improving Risk Analysis Approaches Used by the U.S. EPA, Board on Environmental Studies and Toxicology, Division on Earth and Life studies.

NTP (National Toxicology Program), 2015. Handbook for Preparing Report on Carcinogens Monographs July 20, 2015. Available online: https://ntp.niehs.nih.gov/ntp/roc/handbook/roc_handbook_508.pdf

OECD, 2016. *Users' Handbook supplement to the Guidance Document for developing and assessing Adverse Outcome Pathways*, OECD Series on Adverse Outcome Pathways, No. 1, OECD Publishing, Paris. Available online: https://doi.org/10.1787/5jlv1m9d1g32-en

OHAT, 2015. Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. Research Triangle Park, NC: 30 OHAT.

OSHA, 2016. Guidance on data evaluation for weight of evidence determination: application to the 2012 Hazard Communication Standard.

Oxford Dictionaries, 2017. Available from http://www.oxforddictionaries.com/definition/english/ [Accessed: 30 June 2017].

Perkins EJ, Antczak P, Burgoon L, Falciani F, Garcia-Reyero N, Gutsell S, Hodges G, Kienzler A, Knapen D, McBride M and Willett C, 2015. Adverse outcome pathways for regulatory applications: examination of four case studies with different degrees of completeness and scientific confidence. Toxicological Sciences, 148, 14–25.

SCENIHR (Scientific Committee on Emerging and Newly Identified Health Risks), 2012. Memorandum on the use of the scientific literature for human health risk assessment purposes – weighing of evidence and expression of uncertainty, 19 March, 2012.

SCHER, SCENIHR, SCCS, 2013. Making risk assessment more relevant for risk management. Scientific Committee on Health and Environmental Risks, Scientific Committee on Emerging and Newly Identified Health Risks, and Scientific Committee on Consumer Safety. March 2013.

Small MJ, 2008. Methods for assessing uncertainty in fundamental assumptions and associated models for cancer risk assessment. Risk Analysis, 28, 1289–1308. https://doi.org/10.1111/j.1539-6924.2008.01134.x

Staples C, Mihaich E, Carbone J, Woodburn K and Klecka G, 2004. A weight of evidence analysis of the chronic ecotoxicity of nonylphenol ethoxylates, nonylphenol ether carboxylates, and nonylphenol. Human and Ecological Risk Assessment, 10, 999–1017.

Suter II GW, 2016. Weight of evidence in ecological assessment. Document EPA/100/R16/001. Risk Assessment Forum, U.S. Environmental Protection Agency Washington, DC 20460.

Suter GW 2nd and Cormier SM, 2011. Why and how to combine evidence in environmental assessments: weighing evidence and building cases. Science of the Total Environment, 409, 1406–1417. https://doi.org/10.1016/j.scitotenv.2010.12.029. Epub 2011 Jan 31.

Sutton AJ and Abrams KR, 2001. Bayesian methods in meta-analysis and evidence synthesis. Statistical Methods in Medical Research 10, 277–303.

Thayer KA and Schünemann HJ, 2016. Using GRADE to respond to health questions with different levels of urgency. Environment International, 92–93, 585–589.

Turner RM, Spiegelhalter DJ, Gordon C, Smith S and Thompson SG, 2009. Bias modelling in evidence synthesis. Journal of the Royal Statistical Society. Series A, 172, 21–47.

US EPA (US Environmental Protection Agency), 1998. *Guidelines for Ecological Risk Assessment*. U.S. Environmental Protection Agency, Washington DC.

US EPA (US Environmental Protection Agency), 2000. *Risk Characterization Handbook*. Science Policy Council, US EPA, Washington DC.

Vermeire T, Aldenberg T, Buist H, Escher S, Mangelsdorf I, Pauné E, Rorije E and Kroese D, 2013. OSIRIS, a quest for proof of principle for integrated testing strategies of chemicals for four human health endpoints. Regulatory Toxicology and Pharmacology 67, 136–145. https://doi.org/10.1016/j.yrtph.2013.01.007

WCRF/AICR, 2007. Second expert report - food, nutrition, physical activity and the prevention of cancer: a global perspective systematic literature review - specification manual.

Weed D, 2005. Weight of evidence: a review of concept and methods. Risk Analysis, 25, 1545–1557.

WHO (World Health Organization), 2009. Food Safety. Project to update the principles and methods for the assessment of chemicals in food. Principles and methods for the risk assessment of chemicals in food. EHC 240. ISBN 978 92 4 157240 8.

Wittwehr C, Aladjov H, Ankley H, Byrne HJ and de Knecht HJ, 2016. How adverse outcome pathways can aid the development and use of computational prediction models for regulatory toxicology. Toxicological Sciences, 155, 326–336.

Glossary and Abbreviations

Assembling the evidence	The first of three basic steps of weight of evidence assessment, as proposed in this guidance. Includes identification of potentially relevant evidence, selection of evidence to include in the weight of evidence assessment, and grouping the evidence into lines of evidence (see Sections 2.4 and 4.2)
Best professional judgement	A category of weight of evidence assessment methods involving qualitative listing and qualitative integration of multiple pieces or lines of evidence (see Section 3.1).
Case-specific assessment	Case-specific assessments, where there is no pre-specified procedure and assessors need to choose and apply weight of evidence approaches on a case-by-case basis (see also EFSA, 2016a–c and Section 4).
Causal criteria	A category of weight of evidence assessment methods based on criteria for determining cause and effect relationships (see Section 3.1).
Complementary line of evidence	A line of evidence which can only answer a question or subquestion when it is combined with other line(s) of evidence (see Section 2.1).
Conceptual model	Defined by EFSA (2016b) in the context of environmental risk assessment as 'Step of the environmental risk assessment problem formulation phase describing and modelling scenarios and pathways on how the use of a regulated product may cause harm to a specific protection goal'. A form of conceptual framework, which is defined by Prometheus (EFSA, 2015a,b) as 'The context of the assessment; all sub-question(s) that must be answered; and how they combine in the overall assessment'. In the present guidance, conceptual model refers to a qualitative description or diagram showing how pieces and lines of evidence combine to answer a question or subquestion, as well as any relationships or dependencies between the pieces and lines of evidence. The conceptual model could be presented as, for example, a flow chart or list of logical steps (see Section 4.4).
Consistency	The extent to which the contributions of different pieces or lines of evidence to answering the specified question are compatible (see Section 2.5).
Emergency assessment	Emergency procedures, where the choice of approach is constrained by unusually severe limitations on time and resources (see also EFSA, 2016a–c and Section 4).
Estimate	A calculation or judgement of the approximate value, number, quantity, or extent of something (adapted from Oxford Dictionaries, 2017). Some weight of evidence questions refer to estimates, while others refer to hypotheses (see Section 2.1).
Evidence	Information that is relevant for assessing the answer to a specified question. In PROMETHEUS, a piece of evidence for an assessment is defined as data (information) that is deemed *relevant* for the specific objectives of the assessment (EFSA, 2015b). In this Guidance, this is expanded to all *potentially relevant* information, i.e. all evidence identified by the initial search process, to recognise that the assessment of relevance in the search process is necessarily a preliminary one (e.g. based on keywords and titles alone). 'Evidence' can refer to a single piece of potentially relevant information or to multiple pieces (see Section 2.1).
Expert judgement	EFSA (2014a–d) defines an expert as a knowledgeable, skilled or trained person. An expert judgement is a judgement made by an expert about a question or consideration in the domain in which they are expert. Such judgements may be qualitative or quantitative, but should always be careful, reasoned, evidence-based and transparently documented (see Section 4.4).

GRADE	An approach for grading the quality of evidence and the strength of recommendations in environmental and occupational health, proposed and developed by the Grading of Recommendations, Assessment, Development and Evaluation (GRADE) Working Group (see Morgan et al., 2016).
Hypothesis	One type of framing for weight of evidence questions. Defined by Suter (2016) as a proposition proposed to be a potential explanation of a phenomenon or a potential outcome of a phenomenon. Some weight of evidence questions refer to hypotheses, while others refer to estimates (see Section 2.1).
Influence analysis	A study of possible change in the assessment output resulting not just from uncertainties about inputs to the assessment but also from uncertainties about choices made in the assessment (EFSA, 2016a–c). See Section 4.6.
Integrating the evidence	The third of three basic steps of weight of evidence assessment, as proposed in this guidance. Includes developing a conceptual model for integration, assessing the consistency of the evidence, applying the method chosen for integration and developing the weight of evidence conclusion (see Sections 2.4 and 4.4).
Line of evidence	a set of evidence of similar type (see Section 2.1).
Meta-analysis	a statistical analysis that combines the results of multiple scientific studies (see Section 3.1.5).
OHAT	An approach to systematic review and evidence integration for literature-based environmental health science assessments, developed by the NTP Office of Health Assessment and Translation (OHAT) (see Rooney et al. 2014).
Piece of evidence	A broad term used to refer to distinct elements of evidence that may be combined to form a line of evidence, e.g. a single study, expert judgement or experience, a model, or even a single observation (see Section 2.1).
Probability	Defined depending on philosophical perspective: (1) the frequency with which samples arise within a specified range or for a specified category; (2) quantification of uncertainty as degree of belief regarding the likelihood of a particular range or category (EFSA, 2016a–c). The latter perspective is implied when probability is used in a weight of evidence assessment to express relative support for possible answers (see Sections 2.3 and 2.6).
Problem formulation	In the present guidance, problem formulation refers to the process of clarifying the questions posed by the Terms of Reference, deciding whether and how to subdivide them, and deciding whether they require weight of evidence assessment (Section 2.2).
Qualitative assessment	An assessment performed or expressed using words, categories or labels (see Section 4.1 in EFSA, 2016a).
Quantification	A category of weight of evidence assessment methods defined by Linkov et al. (2009) as comprising formal decision analysis and statistical methods (see Section 3.1). Would also include probabilistic reasoning.
Quantitative assessment	An assessment performed or expressed using a numerical scale (see Section 4.1 in EFSA, 2016a–c).
Rating	A category of weight of evidence assessment methods for weighing and/or integration of evidence based on qualitative logic models, ranks, scores and empirical models (see Section 3.1).
Refinement	one or more changes to an initial assessment, made with the aim of reducing uncertainty in the answer to a question. Sometimes done as part of a 'tiered approach' to risk or benefit assessment.
Relative support	An expression of the extent to which evidence supports one possible answer to a weight of evidence question, relative to other possible answers. Can be expressed qualitatively or quantitatively (see Section 2.3). Quantitative expression can be in terms of probability (see Section 2.6).

Relevance	The contribution a piece or line of evidence would make to answer a specified question, if the information comprising the line of evidence was fully reliable. In other words, how close is the quantity, characteristic or event that the evidence represents to the quantity, characteristic or event that is required in the assessment. This includes biological relevance (EFSA Scientific Committee, 2017) as well as relevance based on other considerations, e.g. temporal, spatial, chemical, etc. (see Section 2.5).
Reliability	Defined in this guidance as the extent to which the information comprising a piece or line of evidence is correct, i.e. how closely it represents the quantity, characteristic or event that it refers to. This includes both accuracy (degree of systematic error or bias) and precision (degree of random error) (see Section 2.5).
Sensitivity analysis	A study of how the variation in the outputs of a model can be attributed to, qualitatively or quantitatively, different sources of uncertainty or variability. Implemented by observing how model output changes when model inputs are changed in a structured way (EFSA, 2016a–c). See Section 4.6.
Standalone line of evidence	A line of evidence which offers an answer to a question or subquestion without needing to be combined with other lines of evidence (see Section 2.1).
Standardised assessment procedures	Assessments where the approach to integrating evidence is fully specified in a standardised assessment procedure. They generally include standardised elements that are assumed to provide adequate cover for uncertainty (EFSA, 2016a–c). See also Section 4.
Uncertainty analysis	A collective term for the processes used to identify, characterise, explain and account for sources of uncertainty (EFSA, 2016a–c). See Sections 2.6 and 4.5.
Uncertainty	A general term referring to all types of limitations in available knowledge that affect the range and probability of possible answers to an assessment question (EFSA, 2016a–c). See Section 2.6.
Variability	Heterogeneity of values over time, space or different members of a population, including stochastic variability and controllable variability (EFSA, 2016a–c). See Section 2.7.
Weighing the evidence	The second of three basic steps of weight of evidence assessment, as proposed in this guidance. Includes deciding what considerations are relevant for weighing the evidence, deciding on the methods to be used, and applying those methods to weigh the evidence (see Sections 2.4 and 4.3).
Weighing	In this guidance, weighing refers to the process of assessing the contribution of evidence to answering a weight of evidence question. The basic considerations to be weighed are identified in this guidance as reliability, relevance and consistency of the evidence (see Section 2.5).
Weight of evidence assessment	A process in which evidence is integrated to determine the relative support for possible answers to a scientific question (see Section 2.1).
Weight of evidence conclusion	the outcome of a weight of evidence assessment, expressed in terms of relative support for possible answers to the weight of evidence question (see Section 2.3).
Weight of evidence question	A question addressed by a weight of evidence assessment. This may be the overall scientific question for an assessment, or a subquestion that contributes to answering the overall question (see Section 2.2). Weight of evidence questions may be framed in terms of hypotheses (which are often qualitative) or estimates (quantitative).
Weight of evidence	The extent to which evidence supports one or more possible answers to a scientific question. Hence 'weight of evidence methods' and 'weight of evidence approach' refer to ways of assessing relative support for possible answers (see Section 2.3)

AICR	American Institute for Cancer Research
AMU	Assessment Methodological unit
ECHA	European Chemicals Agency
FAO	Food and Agriculture Organization of the United Nations
GRADE	Grading of Recommendations, Assessment, Development and Evaluation
IARC	International Agency for Research on Cancer
IPCS	International Programme on Chemical Safety
OECD	Organisation for Economic Co-operation and Development
OHAT	Office of Health Assessment and Translation
PROMETHEUS	promoting methods for evidence use in scientific assessments
SC	Scientific Committee
SCCS	Scientific Committee on Health and Environmental Risks
SCENIHR	Scientific Committee on Emerging and Newly Identified Health Risks
SCHER	Scientific Committee on Consumer Safety
SR	systematic review
US EPA	US Environmental Protection Agency
WCRF	World Cancer Research Fund
WG	Working Group
WHO	World Health Organization

Annex A – Illustration of the proposed approach to assess the Weight of evidence: problem formulation, hierarchy of questions and mapping lines of evidence for chemical risk assessment

Annex A provides examples of the application and reporting of the weight of evidence approaches used in assessing chemical risks to human health and the environment in the context of EFSA's mandate. In the introductory part, general principles of conducting chemical risk assessments are presented, and a decision tree provided to demonstrate data availability and how it drives the applied approaches. This is followed by examples demonstrating different situations with regard to data availability in human health risk assessment and an example of ecological risk assessment. In all cases, specific questions are addressed.

A.1. Problem formulation

Problem formulation refers to framing of the scientific question(s) that are generally posed by a decision maker or a stakeholder. It is important to ensure that the question(s) fully encompass the issue(s) that need to be addressed, and are clear and agreed prior to the start of the assessment (EFSA, 2015c). The WHO (2009) has regarded problem formulation as an iterative process involving risk assessors and risk managers that determines the need for and the extent of a risk assessment. According to the US EPA (2007), problem formulation is 'a systematic planning step to identify the major factors to be considered in a particular assessment in relation to preliminary hypotheses with regards to hazard assessment (i.e. likelihood and severity of adverse effects which might occur or have occurred) and exposure assessment (i.e. likelihood and significance of exposure)'.

In a food safety context, a typical problem formulation may include a description of the intended application (e.g. a food additive) and the commodities involved; issues expected to be affected (e.g. human health), and potential consequences; consumer perception of the hazards or risks; and distribution of possible risks among different segments of the population. The desired outcomes of problem formulation are in the form of clear questions that need answering, identification of the resources and the timeframe that would be needed for the assessment. In relation to EFSA's work, problem formulation is often defined in the Terms of Reference (ToR) provided by risk managers from the European Commission or a Member State. Data requirements for premarket authorisation of a substance are defined in a number of regulation and guidance documents.

With regard to applying weight of evidence in such assessments, it is important to note that problem formulation is a prerequisite and precedes the assessment, including the weighing and the integration of the evidence. In the case of food chemicals, the problem may be too complex to be addressed in a single question, and may need to be divided into more lower tier questions that can be answered directly, and by combining the answers to these, the main question can be addressed. In doing this, a hierarchy of questions is defined. Relevant data/information can be collected, assessed, weighed and integrated into separate lines of evidence that would answer each question at the bottom of the hierarchy. Integration continues upward the question hierarchy until an overall integration can be reached to respond to the main question.

A.2. Chemical risk assessment (human health)

The basic steps of a chemical risk assessment involve a structured way to address the hierarchy of the question(s) for each step of the process, i.e. hazard assessment, exposure assessment and risk characterisation. This gives the assessor a starting point to map the evidence needed for the assessment in a particular context (e.g. regulated chemicals/products or contaminants for human health or environmental risk assessment).

Human risk assessment of chemicals applied by EFSA to the food and feed safety area may include the following generalised elements for hazard identification and characterisation. For each of those elements, different lines of evidence may be available and can be integrated depending on the purpose of the assessment (e.g. a standard procedure, rapid risk assessment, chemical-specific assessment), and availability of data, time and resources.

Hazard identification

- Genotoxicity: *in vitro* studies such as bacterial Ames test and mammalian micronucleus assay; *in vivo* studies; *in silico* models.

- Toxicokinetics: *in vivo* studies on absorption, distribution, metabolism, excretion (ADME); *in vitro* studies; *in silico* models.
- Toxicity/toxicodynamics: epidemiological and clinical studies; case reports; *in vivo* studies (acute, subchronic, chronic toxicity, carcinogenicity, reproductive/developmental toxicity); *in vitro* studies; *in silico* models.

Hazard characterisation (dose–response relationship to derive health based guidance values (HBGV) or reference values (RVs) for margin of exposure (MOE).

- Toxicokinetics: *in vivo* and/ or *in vitro* studies on absorption, distribution, metabolism, excretion (ADME), *in silico* models (toxicokinetic and/ or physiologically based models).
- Acute toxicity: *in vivo* studies, *in vitro* studies, case reports.
- Subchronic/chronic/carcinogenicity: epidemiological and clinical studies; *in vivo* studies including pathological investigations, clinical chemistry; *in vitro* studies; *in vivo* and/or *in vitro* OMICs studies (transcriptomics, proteomics, metabolomics, exposomics); *in silico* models; read-across extrapolations; default values (e.g. threshold of toxicological concern (TTC)).
- Other studies (as necessary): e.g. studies on reproductive/developmental toxicity, neurotoxicity, immunotoxicity.

Exposure assessment

- Occurrence: concentration of the chemical in food (total diet study, food monitoring, food composition tables, etc.) or other media using results of analytical methods (high-performance liquid chromatography (HPLC), gas chromatography (GC), mass spectrometry (MS), etc.), default values (e.g. maximum residue levels (MRLs), maximum limits, etc.), *in silico* models.
- Food Consumption: (consumption surveys, food consumption and composition databases, e.g. EFSA food consumption and composition databases, budget method, volume of production, default value), *in silico* models.

Risk characterisation

- Compounds with threshold effects: comparison of health-based guidance value (e.g. acceptable daily intake (ADI), tolerable daily intake (TDI)) with exposure estimates, MOE approach for chemicals with insufficient data to establish a TDI.
- Compounds with non-threshold effects (e.g. genotoxic and carcinogenic compounds) MOE approach.

The cases described below relate to human risk assessment of regulated products for premarket authorisation, as well as re-evaluation of regulated compounds and contaminants.

A.2.1. Human health risk assessment of regulated products for premarket authorisation

For regulated substances (e.g. food and feed additives, flavourings, food contact materials, pesticides), minimum data requirements are provided in the respective European Commission regulations and relevant guidance documents. The information is required to be submitted in a dossier by any applicant who is seeking authorisation of a substance prior to placing it on the market (see, for example, EFSA, 2012, 2014b). In these cases, problem formulation is often set by the European Commission regulations and guideline documents, defining the hierarchy of the questions for the purpose of the assessment. Where there are data gaps in relation to hazard identification/ characterisation, data/information derived from other methods, such as *in silico* modelling and read-across, may be useful in filling the gaps. Similarly, in the absence of comprehensive data on the occurrence in different food groups, exposure assessment may be based on point estimates.

A.2.2. Re-evaluation of human health risk assessment of existing regulated products

For re-evaluation of regulated substances, available data may include an original dossier, and in some circumstances in addition, historical assessments, and/or published studies in the scientific literature or additional studies provided by applicants that can be used to address each question at different steps of the risk assessment process. The availability of more detailed information would provide more options for applying quantitative probabilistic methodologies.

A.2.3. Human health risk assessment of contaminants

Some contaminants may result from food and feed processing, e.g. acrylamide, PAHs or from natural sources such as toxins of plant, fungal or marine organism origin. Contaminants may also include regulated chemicals that may have been removed from the market, e.g. brominated flame retardants, or other contaminants, e.g., metals, persistent organic pollutants. Although, in the case of chemicals, the hierarchy of the question(s) still follows the structured steps of risk assessment, the nature of the data available may vary enormously. This depends on whether an assessment had been performed previously, or if the contaminant is new and emerging with limited safety data. Consequently, methods to combine evidence may range from basic description of the evidence in data-poor situations to full probabilistic assessment in data-based situations.

A.2.4. A tiered approach to map the level of knowledge and evidence available for human health risk assessment of chemicals

Under the scenarios described above, a tiered approach has been illustrated in Figure A.1 with the aim to map the level of knowledge and the type of evidence/data necessary to address the questions for hazard characterisation of chemicals in human health risk assessment. The tiered approach has been adapted from the WHO, the US EPA and EFSA (US EPA, 2007; Meek et al. 2011; EFSA, 2013). The reader should consider each step of the risk assessment independently since the level of knowledge can be any combination of data-poor and data-based situation for the exposure and/or hazard identification/ characterisation and consequently risk characterisation.

A.2.4.1. Weighing evidence for human risk assessment of chemicals

Once, in the problem formulation, the questions for data-poor and data-rich situations have been defined, and available data sources and data gaps have been identified, a hierarchy of questions can be developed in relation to assembling, weighing and integrating the available evidence.

A.2.4.2. Hazard identification and characterisation

The first step involves searching for and selecting the evidence that is relevant for answering the question(s) at hand and deciding if and how to group it/them into lines of evidence as needed. The second step involves detailed evaluation and weighing of the assembled evidence. In the third step, the evidence is integrated to arrive at conclusions. Each weight of evidence assessment is associated with specific uncertainties that contribute to the overall uncertainty assessment (see Section 2.6 below).

Figure A.1 provides a decision tree which summarises each step of the weight of evidence analysis for hazard identification and hazard characterisation of chemicals in food in data rich and data poor situations. These examples of weight of evidence assessment for hazard characterisation of a chemical in these two data situations are given since these represent both ends of the spectrum EFSA Panels may encounter, when dealing with human health risk assessment of chemicals. In the context of authorisation of regulated compounds, data needs are determined by the relevant guidance and may not cover all endpoints listed, e.g. derivation of a reference point (point of departure) and a health-based guidance value. In some cases, data gaps for specific endpoints may be encountered. In such cases, empirical data can be combined with estimates generated from *in silico* tools.

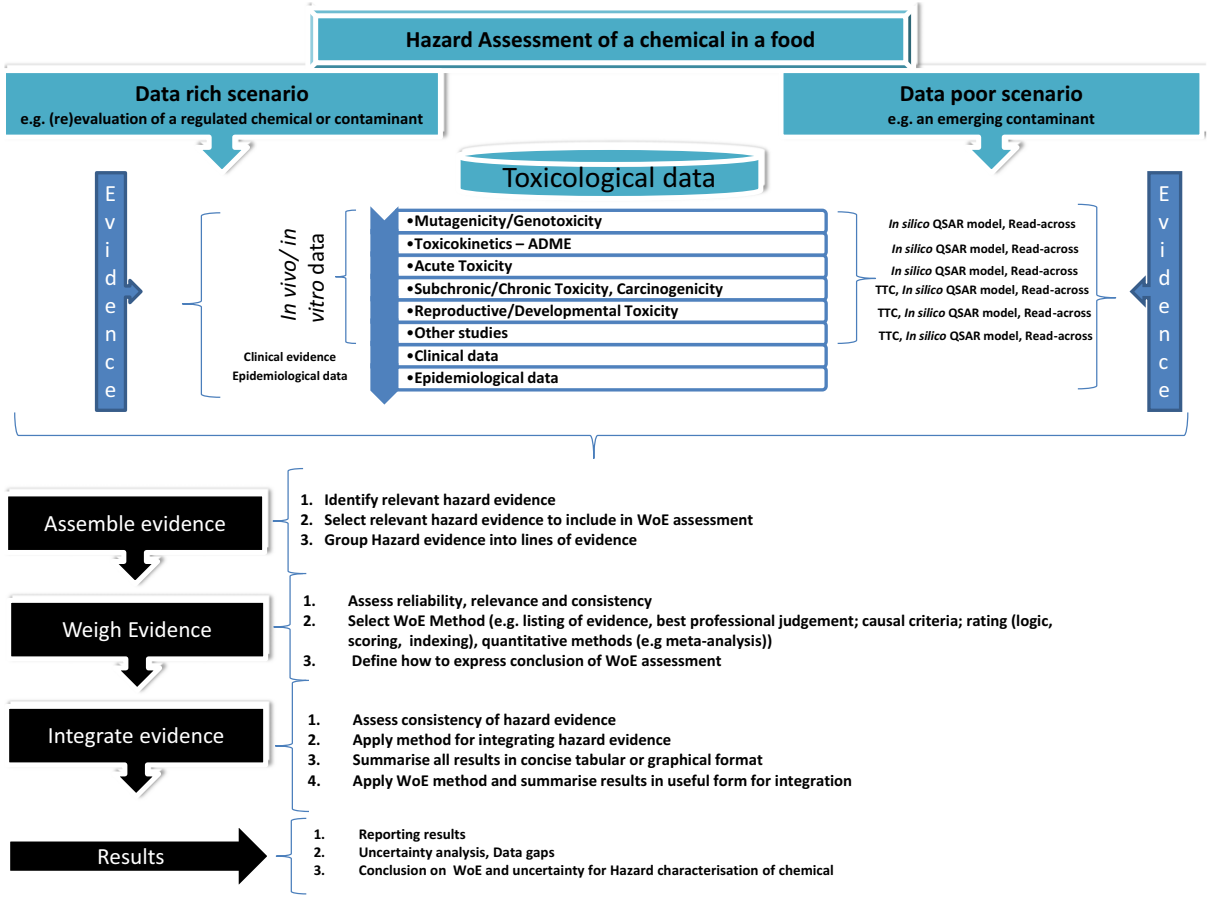

Figure A.1: Generic decision tree for hazard identification and hazard characterisation of chemicals in food in data rich and data poor situations

A.2.4.3. Hazard identification of an emerging contaminant for human health

The example presented below describes the use of weight of evidence in emerging contaminants. They can be anthropogenic contaminants, e.g. brominated contaminants or result from food and feed processing, e.g. newly identified Maillard reaction products or even from natural sources such as toxins of plant, fungal or marine organism origin.

Table A.1: Optional tabular format for summarising weight of evidence assessment for an emerging contaminant

Question: Hazard identification of an emerging contaminant		
Assemble the evidence	Select evidence	No toxicity data available: use read-across from already-tested similar compounds, *in silico* tools (QSAR) to predict toxicity
	Lines of Evidence	Identify lines of evidence for potential effect(s) from the presence of a structural alert or QSAR models, read-across from similar compounds
Weigh the evidence	Methods	Evaluate the reliability, relevance and consistency of the QSAR models. This can include weighing model results on a statistical basis (e.g. likelihood of a compound with a structural alert to express (a) toxic property(ies))
	Results	Toxicity value for each line of evidence, with associated assessment of reliability (e.g. through the applicability domain of the models used)
Integrate the evidence	Methods	If the estimates from the different models converge, the level of uncertainty regarding the toxic property(ies) can be evaluated (e.g. through the applicability domain of the models used). If the estimates do not converge, further modelling for the toxic property(ies) could be undertaken to evaluate whether the results can be improved
	Results	Integrated the toxicity value and uncertainty factor to derive a health based guidance value for the emerging contaminant: Summary Table

Preferably use multiple tools for the assessment, for instance, multiple QSAR models. Different *in silico* tools will have different levels of transparency, and may be more or less relevant for the target compound. Preference should therefore be given to the model(s) that also provide an assessment of their applicability domain(s). A general strategy for weighing the evidence from different tools should consider the following criteria:

1) Does the chemical possess a (structural/functional) feature that indicates potential for toxicity (e.g. the presence of a structural alert)?
2) Are there other factors that may negate this feature (e.g. an exception rule to the structural alert, or other influencing factors such as the lack of systemic absorption)?
3) Are there similar compounds with the same feature as the target compound (again considering any counteracting factors)?
    a) Do the similar compounds have the expected toxicity, as suggested by the presence of the alert?
    b) If there are no similar compounds with the relevant toxicity feature, are there other structurally or functionally similar compounds that have been tested?
4) Do the results of the QSAR models support the focussed studies listed above?

Note that a compound may still be toxic even if it is not associated with any known alert or not toxic even if it contains a structural alert for toxicity.

In straight forward cases, there is generally an agreement between the presence of a toxicity alert, read-across from similar compounds and predicted toxicity by QSAR model(s). In more complex situations, however, some alerts may conflict with the evidence from different methods, e.g. similar compounds used in a read-across may have conflicting values, or the output from different QSAR models may show disagreement. The use of more sophisticated *in silico* tools (such as the OECD toolbox, VEGA, etc.) may help the expert in identifying any toxicity alerts in the target compound, and/or the presence of the alerts in other similar compounds, as well as providing a measure of the uncertainty associated with each alert and result of read-across/QSAR modelling.

A.2.4.4. Example of the use of non-testing methods within a weight of evidence framework

The example presented below describes the use of weight of evidence when information is derived from non-testing methods (NTM), such as *in silico* models and read-across. Two *in silico* platforms – VEGA (www.vega-qsar.eu) and T.E.S.T. (www.epa.gov/nrmrl/std/qsar/qsar.html#TEST) – have been used to estimate mutagenicity (as assessed through the bacterial reverse mutation test) of the target compound. A read-across software, ToxRead (www.toxread.eu), has also been used to provide additional support for the results where necessary.

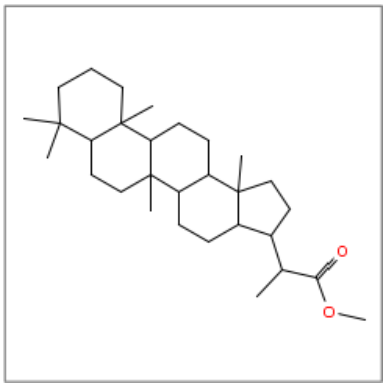

Figure A.2: shows chemical structure of the compound used in *in silico* assessment of toxicity

Table A.2: Summary of the results obtained by different non-testing methods

Software	Model/method	Results	Applicability Domain Index
T.E.S.T.	Consensus method Hierarchical method FDA method Nearest neighbour method	Mutagenicity negative Mutagenicity negative Mutagenicity negative Mutagenicity negative	Internally checked
VEGA	Consensus model CAESAR(a) SARpy[3] ISS[3] KNN[3]	Non-mutagenic Mutagenic Non-mutagenic Non-mutagenic Non-mutagenic	0.719 0.853 0.786 0.819
ToxRead	Read-across	Non-mutagenic	Not available

(a): These are the names of models as found in VEGA (www.vega-qsar.eu).

As shown in Table A.2, all models predicted the substance as non-mutagenic, except the CAESAR model within the VEGA platform. Indeed, the target compound does not have a structural rule for mutagenicity. On the contrary, seven rules associated with the lack of effect have been reported by SARpy within the VEGA platform. Most of the similar compounds in this case are also not-mutagenic, and the majority of the QSAR models used also indicate non-mutagenicity. Thus, the three main elements in the evaluation using non-testing methods (structural rules, read-across from similar compounds and QSAR models) agree on the lack of mutagenic effect. These results are based on experimental values for structurally similar compounds and on the results of the QSAR models, using both rule- and statistically based models. Since no rule has been identified for mutagenicity, reasoning for the mechanism is not applicable. The read-across tool (ToxRead) confirms that the substance should not be regarded mutagenic.

However, two indications for mutagenicity need further explaining. First, the CAESAR model indicates mutagenicity, and there is a structurally similar compound (CAS 1250-95-9) that is reported mutagenic by both VEGA and T.E.S.T. The applicability domain of the CAESAR model for this compound is nevertheless weak, and the software indicates a particular warning highlighting this aspect. This, coupled with the fact that similar compounds in the training sets are also non-mutagenic, suggests that the model prediction for mutagenicity of the target compound is not correct. The VEGA software clearly indicates that the very likely reason of the mutagenic effect of the structurally similar compound is the presence of an epoxide moiety. However, this moiety is not present in the target

compound, and therefore, this conflicting prediction can be disregarded. This example shows proper use of the criterion about relevance of the line of evidence. In this case, mutagenicity of similar substances with moieties that are not present in the target compound makes them non-relevant to the target compound.

The reliability of the other lines of evidence is addressed in this case through the use of the measurement of the applicability domain, as in Table A.2. In this way, the user can integrate the different lines of evidence, keeping into account their relevance, reliability, and consistency.

The read-across software (ToxRead) provides further support to this conclusion. Most of the closely related substances (Figure A.3) show a lack of mutagenicity, whereas one similar compound which is mutagenic (indicated in red circle) is in fact the same as discussed above.

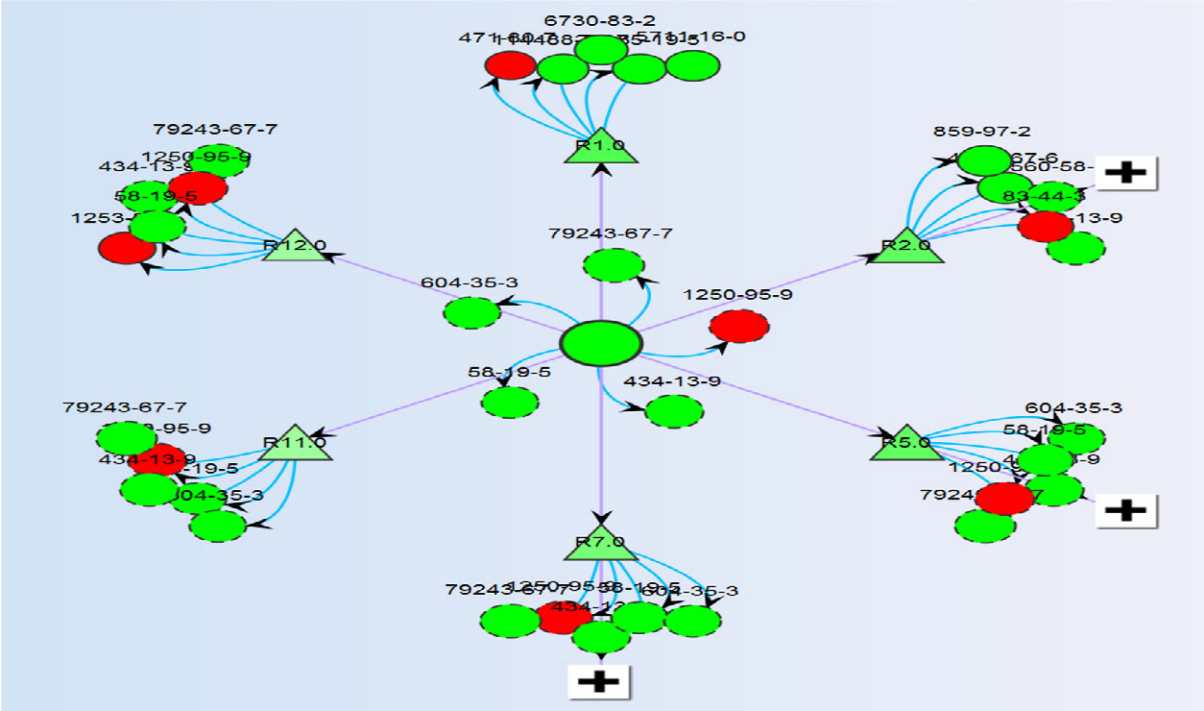

Figure A.3: ToxRead chart (www.toxread.eu). The numbers refer to CAS identifiers. Straight arrows link the target chemical to rules, while curved arrows link to chemicals

The overall conclusion that can be drawn from this exercise is that the substance in question is very likely not mutagenic (90% of probability).

This example illustrates how different tools can be used to derive toxicity estimates for a given compound in the absence of experimental data. Some of these tools may be rule-based (e.g. VEGA ISS), statistically based (e.g. T.E.S.T) and tools which offer both these approaches, and thus go beyond simple prediction and provide elements for explicit reasoning. For example, some programmes (e.g. VEGA) provide necessary details that facilitate the reasoning, and thus increase the acceptance of the results. The read-across tool, ToxRead, provides a weight of evidence program that can integrate results from QSAR and read-across. As shown in the example above, any contradictory results from the different methods need to be explored further before considering or disregarding them. Therefore, such tools must not be used as a 'black box', and the final assessment must be carried out by an expert.

Table A.3: Optional tabular format for summarising weight of evidence assessment of an emerging contaminant

Question		*Hazard identification of an emerging contaminant*
Assemble the evidence	Select evidence	Nine QSAR models from two *in silico* platforms and a program for read–across were used to estimate mutagenicity potential (as assessed through bacterial reverse mutation test) of the target compound
	Lines of evidence	Except two, all estimates indicated the compound to be non-mutagenic. The exception was the QSAR model CAESAR within VEGA platform that predicted the compound as mutagenic, and the read-across programme ToxRead that showed one out of five similar compounds to be mutagenic
Weigh the evidence	Methods	VEGA provides a quantitative measurement of reliability and values higher than 0.8 ADI are considered more reliable. T.E.S.T. applies a filter to eliminate not reliable predictions. The results obtained from these platforms in this case are therefore reliable. ToxRead indicates the alerts associated with the effect and similar compounds. In case of chemicals with the toxicity value conflicting with the rule, the user should check if there are rules present only in the similar compound and not in the target, explaining the conflicting toxicity value. This is useful to evaluate the relevance of the lines of evidence, disregarding those that are not relevant
	Results	T.E.S.T. results consistently indicated non-mutagenicity. The VEGA models called SARpy and KNN showed higher indices for reliability, also predicted non-mutagenicity. The CAESAR and ISS models within the VEGA models showed relatively lower reliability. ToxRead results show that most of the compounds similar to the target compound were not-mutagenic. The only structural rule for mutagenicity found in one similar compound is not present in the target compound, and therefore is not relevant
Integrate the evidence	Methods	The *in silico* estimates have been integrated while considering the reliability and relevance of the individual values, together with the consistency of all the predicted values, to make an informed expert judgement about the probability that the target compound is not-mutagenic
	Results	The large majority of the *in silico* values are in concordance for non-mutagenicity of the target compound. One conflicting estimate is less reliable whereas the other is not relevant to the target compound. Considering all the evidence from this *in silico* assessment, it was concluded by informed expert judgement that the target compound is most likely (about 90% probability) to be non-mutagenic

ADI: Applicability Domain Index.

A.2.4.5. Example of the use of human epidemiological data within a weight of evidence framework: The cadmium example

The CONTAM Panel performed a human risk assessment for cadmium in food in 2009 (Amzal et al., 2009; EFSA, 2009) and its assessment is an example of the use of a weight of evidence (WoE) approach using human epidemiological data to derive a reference point based on benchmark dose limit (BMDL). Cadmium (Cd) is a heavy metal found as an environmental contaminant, both through natural occurrence and from industrial and agricultural sources. Foodstuffs are the main source of cadmium exposure for the non-smoking general population. Cadmium absorption after dietary exposure in humans is relatively low (3–5%), but cadmium is efficiently retained in the kidney and liver in the human body, with a very long biological half-life ranging from 10 to 30 years.

The CONTAM Panel considered human studies relating to urinary cadmium and urinary biomarkers of toxicity for kidney toxicity (*N*-acetyl-f3-glucosaminidase, beta–microglobulinuria (B2-M), alpha-1-microglobulinuria, urinary retinol-binding protein, proteinuria) and bone effects (bone mineral density, alkaline phosphatase activity, serum calcium, parathyroid hormone) using a systematic review of the literature. Based on the literature availability at the time, expert judgement and previous international risk assessments (JECFA, ATSDR), the CONTAM Panel concluded that B2-M was the most reliable, relevant and consistent urine biomarker for Cd-induced renal tubular toxicity with 35 studies reporting continuous variables as preferable for benchmark dose modelling.

A Bayesian meta-analysis and hierarchical modelling was performed to build an overall dose-effect relationship accounting for interstudy heterogeneity and for inter-individual variability of dose and

effect. Subsequently, a BMDL was evaluated using a hybrid approach for various cut-offs. As a lower and more protective cut-off level, the Panel proposed a biological cut-off for B2-M of 300 μg/g creatinine from expert judgement and clinical evidence that exceeding such a threshold is associated with an accelerated age-related decline renal function together with increased mortality (Amzal et al., 2009). The CONTAM Panel selected an overall group-based BMDL5 of 4 μg cadmium/g creatinine. The use of 300 g B2-M/g creatinine as critical effect of cadmium exposure to base the risk assessment leads to a possible overestimation of the risk, but is protective of the most sensitive groups of the population. A summary of the WoE assessment for deriving a BMDL for cadmium is presented in Table A.4.

Table A.4: Summary of the weight of evidence assessment for the derivation of a human reference point (benchmark dose limit) for cadmium in food

Question		**Deriving a reference point (benchmark dose limit) for cadmium in humans (hazard characterisation)**
Assemble the evidence	Select evidence	Systematic review of human studies relating urinary cadmium and excretion of biomarkers of toxicity Select relevant papers: biomarkers of kidney and bone effects with continuous outcome to include in WoE assessment
	Lines of evidence	Urinary cadmium and renal effects LOE 1: *N*-acetyl-f3-glucosaminidase (NAG) LOE 2: beta-microglobulinuria (B2-M) LOE 3: alpha-1-microglobulinuria LOE 4:urinary retinol-binding protein (RBP) LOE 5: proteinuria Urinary cadmium and bone effects LOE 1: bone mineral density (bone MD) LOE 2: Alkaline phosphatase activity (bALP) LOE 3: serum calcium LOE 4: parathyroid hormone (PTH)
Weigh the evidence	Methods	1. Assess reliability, relevance and consistency by expert judgement 2. Select WoE Method: Quantitative method as meta-analysis using Bayesian hierarchical mixed effect model to build dose-response relationship between urinary cadmium and urinary B2-M 3. Conclusion for WoE Assessment: Dose-effect relationship accounting for inter-study heterogeneity and for inter-individual variability of urinary cadmium and excretion of B2-M selected for the modelling
	Results	Overall 35 studies B2-M studies (LOE 1) were selected as most relevant, consistent and reliable including previous meta-analysis and assessments from ATSDR, WHO Urinary cadmium and B2-M selected for meta-analysis based on biological relevance of the biomarker for cadmium toxicity, reliability and consistency of the human database
Integrate the evidence	Methods	Meta-analysis of the dose-effect relationship between urinary cadmium and B2-M accounting for interstudy heterogeneity and for interindividual variability of dose and effect using Bayesian hierarchical mixed effect model Expert judgement was used to select biologically relevant cut off values for urinary B2-M reflecting renal tubular damage (300 μg B2-M/g creatinine). Exceeding such a cut off value has been associated with an accelerated age-related decline of renal function together with increased mortality Derive the reference point based on benchmark dose limit (BMDL5) modelling for urinary B2-M using hybrid approach and urinary cadmium
	Results	The meta-analysis and dose–response modelling based on B2-M as a marker of tubular effect, identified an overall group based BMDL5 for a 5% increase of the prevalence of elevated B2-M of 4 μg cadmium/g creatinine. This is selected as a reference point

WoE: weight of evidence; LOE: line of evidence.

A.2.4.6. Example of the use of weight of evidence framework in derivation of an acceptable daily intake (ADI) for a regulated chemical

EFSA may derive an acceptable daily intake (ADI) for regulated products that are likely to be present in food or feed. The ADI is an estimate of the amount of a chemical that can be consumed on a daily basis over a lifetime without appreciable health risk. Within the context of EFSA risk assessments, ADIs are derived for food and feed additives, and pesticide residues. An applicant wishing to market a regulated product is required to demonstrate its safety by providing data from relevant toxicity studies, from which an ADI can be derived. For example, the EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS) uses a tiered approach with increasing data requirement in higher tiers reflecting greater potential risk (EFSA, 2012, 2014d). The data set needs to address toxicokinetics, genotoxicity, subchronic toxicity, chronic toxicity, carcinogenicity, reproductive toxicity and developmental toxicity. These studies may be performed in several species such as rat, mouse, dog, rabbit and each type of study per species constitutes a line of evidence (LOE) (Table A.5).

For each line of evidence, effect data from the different species may be compared qualitatively to identify the no-observed-adverse-effect-level (NOAEL) for the most sensitive species and endpoint taking into account biological relevance, reliability and consistency of the data. Alternatively, benchmark dose modelling could be performed on the results from each study using model averaging according to the guidance of EFSA scientific Committee (EFSA Scientific Committee, 2017). Quantitative comparison of the results of the BMDL for each study and LOE can then be carried out to determine a BMDL for the most sensitive species and endpoint taking into account biological relevance, reliability and consistency of the data.

An ADI is finally derived by dividing the NOAEL or BMDL by an appropriate uncertainty factor to account for differences in TK and TD between experimental animals and human (10-fold), and variability among humans (10-fold). An additional uncertainty factor may in some cases be applied to account for severity of the effect or deficiency in the data.

Table A.5: Summary of the weight of evidence assessment for the derivation of an acceptable daily intake (ADI) for a regulated non-genotoxic chemical

Question		**Derivation of an acceptable daily intake (ADI) for a regulated non-genotoxic chemical**
Assemble the evidence	Select evidence	In the context of a regulated compound, toxicity studies that may be used to derive an ADI are illustrated below (this list is not exhaustive). These studies may be performed in several species namely rat, mouse, dog, rabbit, and each type of study per species constitute a line of evidence
	Lines of evidence	Examples include LOE 1: subchronic toxicity study 28 days LOE 2: subchronic toxicity study 90 days LOE 3: developmental toxicity study LOE 4: one generation reproductive toxicity study LOE 5: chronic toxicity studies (e.g. 1 year toxicity study) LOE 6: Two-year carcinogenicity study
Weigh the evidence	Methods	A number of options may be available depending on the specific assessment. Generic examples are illustrated below: 1) Qualitative comparison of each LOE per species to derive a reference point such as a BMDL or a no-observed-adverse-effect-level (NOAEL) for the most sensitive species and endpoint taking into account biological relevance, reliability and consistency of the data 2) Benchmark dose modelling using model averaging according to the guidance of EFSA scientific Committee (EFSA Scientific Committee, 2017). Quantitative comparison of the results of the BMDL for each study and LOE to determine a BMDL for the most sensitive species and endpoint taking into account biological relevance, reliability and consistency of the data

Question		Derivation of an acceptable daily intake (ADI) for a regulated non-genotoxic chemical
		3) Expert judgement to assess whether an extra UF is needed for the severity of the effect or uncertainty in the DB for the NOAEL or the BMDL
	Results	Possible outcomes include: • Reference point (BMDL or NOAEL) value for the regulated compound for the most sensitive species and endpoint BMDL or NOAEL selected as a reference point • The quality of the studies in terms of relevance, reliability and consistency do not allow the derivation of a reference point • No effects were observed at the highest dose tested and there is no need to derive a numerical reference point • Decision as to whether an extra UF should be applied and how large it should be
Integrate the evidence	Methods	Quantitative methods: combination of reference point and default uncertainty factors
	Results	A number of options may be available depending on the specific assessment. Four options are illustrated below: 1) Derivation of the ADI applying the default uncertainty factor of 100 to the NOAEL, taking into account species differences (10-fold) and human variability (10-fold) 2) Derivation of the ADI applying the default uncertainty factor of 100 to the BMDL taking into account species differences (10-fold) and human variability (10-fold) 3) Derivation of the ADI applying the default uncertainty factor of 100 and an extra uncertainty factor for the severity of the effect (e.g. carcinogenicity) or uncertainty in the database to the NOAEL or BMDL 4) Derivation of the ADI is not possible since regulatory requirements are not met and additional toxicity studies are required 5) Derivation of a numerical ADI is not needed since regulatory requirements are met and no adverse effects were observed at the highest dose tested

LoE: line of evidence.

A.3. Chemical risk assessment (Environment)

While environmental risk assessment of chemicals follows basically the same conceptual framework like that for human health, it is not identical. For hazard identification, all available studies are assessed, and the most sensitive are selected. Available data are integrated to derive a guidance value such as a regulatory acceptable concentration. In the risk characterisation step, actual environmental exposure levels are compared with this guidance value.

Figure A.4 provides a decision tree for environmental hazard characterisation of a chemical in data-based and data-poor situations.

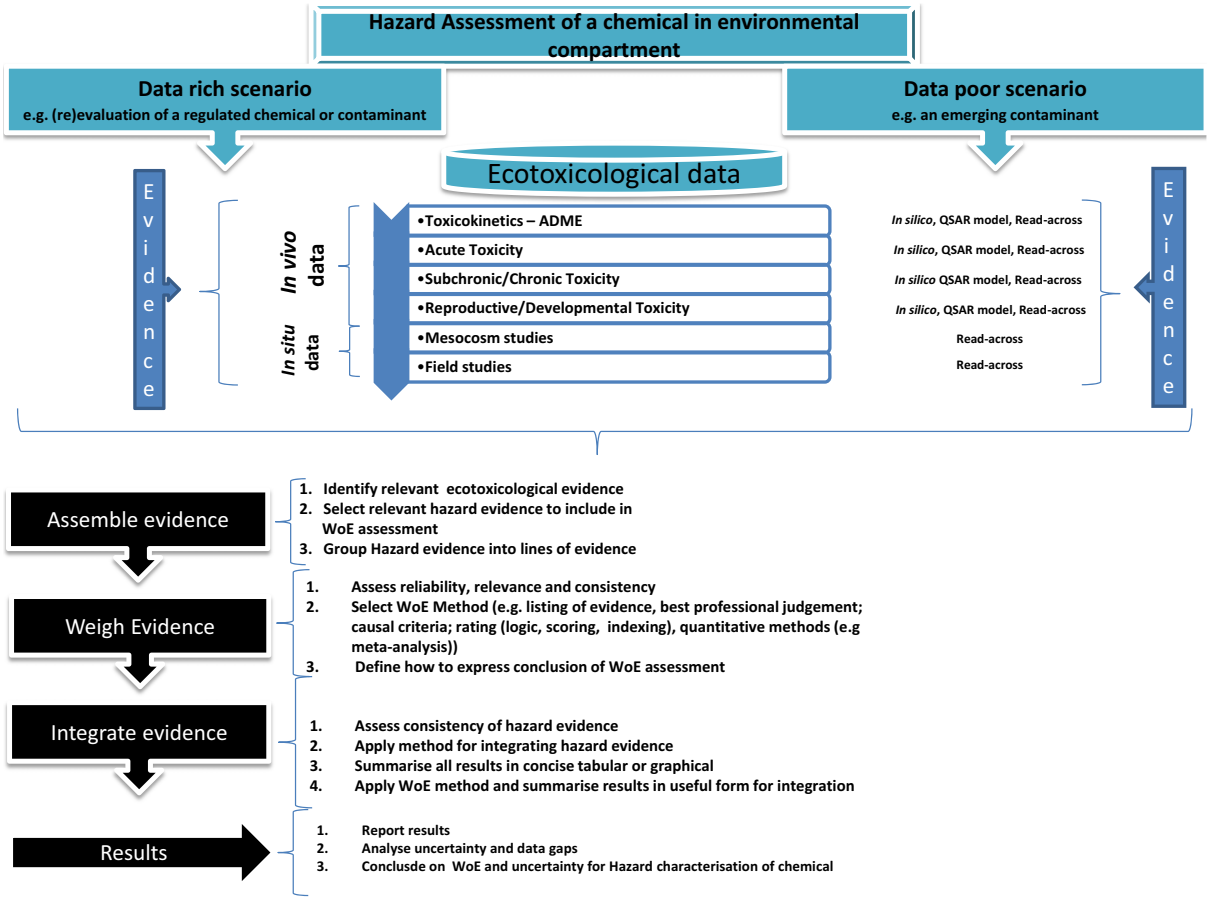

Figure A.4: Decision tree for environmental risk assessment of a chemical in data rich and data poor situations

A.3.1. Ecotoxicological hazard characterisation of a regulated substance

Ecotoxicity studies from the dossier: *in vivo* studies from the dossier (lethality, developmental, reproductive, acute, subchronic and chronic toxicity). Select most sensitive biologically relevant endpoints from a study or all studies when available for a specific group of animals (see note 1). Where more studies are available for the ecotoxicological characterisation than the basic dossier requirements, select most sensitive biologically relevant endpoints by comparing all studies. Use default assessment factor (see note 2) when only dealing with dossier studies, or assess whether there is a need to divert from the default assessment factor on the basis of the available information for assessing the uncertainty and/or variability. Integrate ecotoxicity values and assessment values to derive environmental based guidance value (or regulatory acceptable concentration) for a regulated compound.

The Table below shows an example assessment according to the Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters (EFSA, 2013)

Table A.6: Summary of the weight of evidence assessment for plant protection products for aquatic organisms in edge-of-field surface waters

Question		**Assessing a regulatory acceptable concentration (RAC) for fish for compound Idefix**
Assemble the evidence	Select evidence	Take the ecotoxicological information from the dossier
	Lines of evidence	Acute toxicity values: LOE 1: 5-day LC_{50} study with *Oncorhynchus mykiss* of 12 mg/L LOE 2: 5-day LC_{50} study with *Lepomis macrochirus* of 47 mg/L Chronic toxicity studies: LOE 1: 35-day NOEC study (ELS) with *Oncorhynchus mykiss* of 0.8 mg/L LOE 2: 35-day NOEC study (ELS) with *Lepomis macrochirus* of 0.5 mg/L Standard/default assessment factors: LOE 1: for acute situation assessment factor is 100 LOE 2: for chronic situation assessment factor is 10
Weigh the evidence	Methods	Take lowest value according to the method described in the guidance document and apply appropriate assessment factor. If this results in concern being raised, consider by expert judgement whether to use an alternative value, e.g. the geometric mean (EFSA 2005)
	Results	For the acute situation, the lowest toxicity value is 12 mg/L For the chronic situation the lowest toxicity value will 0.5 mg/L
Integrate the evidence	Methods	Quantitative combination of a point estimate with a default value
	Results	For the acute situation, the RAC is 12/100 = 0.12 mg/L For the chronic situation, the RAC is 0.5/10 = 0.05 mg/L

Note 1. Here available studies are part of a dossier. In case of a new compound/product, the dossier will contain only the ecotoxicological studies required by legislation. This could be one or (in a few cases) two studies. Standard ecotoxicological studies are performed on behalf of, or by, the applicant, which implies that there are also standard endpoints which are in most cases considered as biologically relevant. There are also standard species tested, e.g. *Daphnia magna* as the standard species for crustaceans or the rainbow trout as the standard species for fish. The weight of evidence in those cases has been applied before, choice of representative species or biological relevant endpoints (Commission Regulation (EU) No 283/2013, Commission Regulation (EU) No 284/2013). The uncertainties around the hazard c.q. risk assessment are assumed to be covered by the assessment factor that is used for decision making on whether or not to allow the compound on the market or for going to higher tier risk assessment steps.

Note 2. It is assumed that the assessment factors are only to be used for the uncertainty inherent to the ecotoxicity values and that the uncertainty around the exposure value is included in the assessment of the exposure value, for instance by using the 90th percentile of the exposure distribution.

References

EFSA, 2009. Scientific Opinion of the Panel on Contaminants in the Food Chain on a request from the European Commission on cadmium in food. EFSA Journal 2009; 980, 1–139.

Amzal B, Julin B, Vahter M, Wolk A, Johanson G, Akesson A; Environ Health Perspect, 2009. Population toxicokinetic modeling of cadmium for health risk assessment. 117, 1293–1301. https://doi.org/10.1289/ehp.0800317

EFSA ANS Panel (EFSA Panel on Food Additives and Nutrient Sources added to Food), 2012d. Guidance for submission for food additive evaluations. EFSA Journal 2012;10(7):2760, 60 pp. https://doi.org/10.2903/j.efsa.2012.2760

Meek ME(Bette), Boobis AR, Crofton KM, Heinemeyer G, Raaij MV, Vickers C, 2011. Risk assessment of combined exposure to multiple chemicals: A WHO/IPCS framework. 2011. Regulatory Toxicology and Pharmacology, https://doi.org/10.1016/j.yrtph.2011.03.010

US EPA, 2007. Concepts, Methods and Data Sources for Cumulative Health Risk Assessment of Multiple Chemicals, Exposures and Effects: A Resource Document. EPA/600/R-06/013F. Washington, DC: SEPA. http://cfpub.epa.gov/ncea/risk/recordisplay.cfm?deid=190187

EFSA (European Food Safety Authority), 2013. International Framework Dealing with Human Risk Assessment of Combined Exposure to Multiple Chemicals. EFSA Journal 2013;11(7):3313, 69 pp. https://doi.org/10.2903/j.efsa.2013.3313

EFSA, 2017. EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen KH, More S, Mortensen A, Naegeli H, Noteborn H, Ockleford C, Ricci A, Rychen G, Silano V, Solecki R, Turck D, Aerts M, Bodin L, Davis A, Edler L, Gundert-Remy U, Sand S, Slob W, Bottex B, Abrahantes JC, Marques DC, Kass G and Schlatter JR, 2017. Update: EFSA 2017. Guidance on the use of the benchmark dose approach in risk assessment. EFSA Journal 2017;15(8):4658, 41 pp. https://doi.org/10.2903/j.efsa.2017.4658

EFSA PPR Panel (EFSA Panel on Plant Protection Products and their Residues), 2005. Opinion of the Scientific Panel on Plant protection products and their residues (PPR) on a request from EFSA related to the evaluation of pirimicarb, EFSA Journal 2005;3(8):240, 21 pp. https://doi.org/10.2903/j.efsa.2005.240

Commission Regulation (EU) No 283/2013 of 1 March 2013 setting out the data requirements for active substances, in accordance with the Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of plant protection products on the market. OJ L 93, 3.4.2013, p. 1–84.

Commission Regulation (EU) No 284/2013 of 1 March 2013 setting out the data requirements for plant protection products, in accordance with the Regulation (EC) No 1107/2009 of the European Parliament and of the Council concerning the placing of plant protection products on the market. OJ L 93, 3.4.2013, p. 85–152.

Annex B – Examples of weight of evidence definitions, criteria and outputs from the scientific literature

This Annex contains examples referred to in Section 2 of the Guidance, on three topics:

- Examples of definitions and descriptions of weight of evidence (Table B.1).
- Examples of definitions and descriptions of 'Line of evidence' from the published literature (Table B.2).
- Examples of criteria for weighing evidence from the published literature, mapped onto the three general concepts of reliability, relevance and consistency (Table B.3).

All three tables contain examples from a selection of publications on weight of evidence assessment. They do not comprise a systematic or comprehensive review.

Table B.3 aims to show how various criteria used in the literature relate to the three basic concepts introduced in Section 2.5 of this Guidance: reliability, relevance and consistency. Many publications on weight of evidence approaches specify criteria to be considered when weighing evidence. Some of these criteria are very general, applying to virtually any context. Others are expressed in ways that are more specific to particular problem areas, e.g. the essentiality of key events (Collier et al., 2016). As may be expected, in some cases different publications use different words to express what appear to be similar criteria.

Criteria referring to the quality of studies or data, risk of bias, imprecision and sensitivity are all factors affecting the reliability of a piece of evidence in itself. They are all aspects of the way a study (for example) was conducted or reported, which affect the reliability of the resulting evidence as a correct representation of what actually occurred in the study.

Relevance concerns the relation between evidence and a purpose for which it is being used, i.e. the question it is being used to address. Confounding is a factor affecting relevance, because it concerns the possibility that an observed effect may have been caused by agents or factors other than the one of interest for the question in hand. The same applies to other criteria relating to attribution of an effect, e.g. specificity, temporality, essentiality, experimentation and randomised trials. Criteria such as spatial and temporal representativeness concern the relevance of the conditions in which evidence was generated to the conditions for the question being assessed: for example, if the assessment question refers to the EU as a whole, old data from one EU Member state is less relevant than new data from an EU-wide study (other things being equal).

Some criteria refer to the magnitude or statistical significance of measurements, or of dose-response and other forms of association or correlation (including spatial and temporal correlation). These criteria can contain elements of both reliability and relevance. For example, when an observed effect is large in magnitude and/or shows a consistent trend or association (referred to as 'strength' of evidence by Suter and Cormier, 2011), then it is more likely to be real (reliability of the finding) and more likely to be caused by the agent of interest or more likely to be large enough to be important (relevance to the question). Therefore, in this guidance, 'strength' is considered to contribute to reliability and relevance, and not a separate criterion distinct from reliability and relevance (as in Suter, 2016).

Some other types of criteria can also influence both reliability and relevance. For example, the use of standard methods, the clarity and completeness of reporting and the extent of evaluation and peer review all influence judgements about both the reliability and relevance of evidence.

Many publications include criteria relating to the consistency of evidence. Consistency intrinsically includes the notions of quantity and diversity, because consistency has more weight when seen in a larger body of evidence, and/or when the evidence is of diverse types. Many publications include explicit criteria for consistency, quantity or diversity of evidence, or obviously related criteria such as coherence, reproducibility and replicability. Some of the other criteria in Table B.3 refer particular types of consistency: for example, concordance and analogy both refer to whether there is consistency between the evidence being considered and established knowledge or theory, and hence with the evidence on which the knowledge or theory is based. 'Biological plausibility' usually refers to consistency between data and biological theory or mechanism. Similarly, 'experimental verification' refers to one piece of evidence being supported by another.

Criteria listed in the right hand column of Table B.3 are of a different nature. 'Adequacy' is mentioned by several publications but refers to standards against which reliability, relevance and consistency should be judged, rather than being a separate criterion. 'Validity' also implies a comparison with some standard. In the context of weight of evidence, validity could refer to meeting standards for reliability (e.g. required levels of precision and accuracy), but it can also refer to meeting

standards for relevance (e.g. appropriate test subjects and route of exposure) or even consistency. 'Plausibility' as a general term (distinct from 'biological plausibility', see above) is not really a separate criterion, rather it is a function of the three basic criteria identified in the guidance: relevance, reliability and consistency all contribute to plausibility. It can be regarded as synonymous with the 'relative support' for a given estimate or hypothesis, e.g. regarding a stressor or mechanism. SCENIHR (2012) propose that assessors should 'identify uncertainties in the judgement used' in the weight of evidence process: this refers to the need to consider uncertainties affecting the assessment of reliability, relevance and consistency, rather than being a separate criterion.

Table B.1: Examples of definitions and descriptions of weight of evidence

Publication	Definitions or descriptions given for weight of evidence
Ågerstrand and Beronius (2016)	'In general terms, weight of evidence and systematic review are processes of summarising, synthesising and interpreting a body of evidence to draw conclusions . . . these processes differ from the traditional method for risk assessment by promoting the use and integration of information from all the available evidence instead of focusing on a single study'
ANSES (2016)	Defines weight of evidence as 'the structured synthesis of lines of evidence, possibly of varying quality, to determine the extent of support for hypotheses'
Beronius et al. (2014)	States that 'The meaning of weight of evidence intended here is the collective summary and evaluation of all existing evidence after a certain "weight" has been attributed to individual studies, e.g. by evaluating reliability and relevance'
Collier et al. (2016)	Describes weight of evidence as 'a term used in multiple disciplines to generally mean a family of approaches to assess multiple lines of evidence in support of (or against) a particular hypothesis, although (it) tends to be used inconsistently and vaguely across disciplines'
ECHA (2015a) [Guidance on the Application of the CLP Criteria]	'A weight of evidence determination means that all available information bearing on the determination of hazard is considered together'
ECHA (2015b) [Guidance on the Biocidal Products Regulation]	'A weight of evidence assessment involves the consideration of all data that is available and may be relevant to reproductive toxicity'
ECHA (2010)	'One definition for weight of evidence is: "the process of considering the strengths and weaknesses of various pieces of information in reaching and supporting a conclusion concerning a property of the substance"'
EFSA (2013) [PPR Aquatic Ecol RA guidance doc]	States that the 'process of combining available lines of evidence to form an integrated conclusion or risk characterisation is frequently referred to as weight-of-evidence assessment. This term reflects the principle that the contribution of each line of evidence should be considered in proportion to its weight'
EPA (2003)	Describes weight of evidence as an 'approach (which) considers all relevant information in an integrative assessment that takes into account the kinds of evidence available, the quality and quantity of the evidence, the strengths and limitations associated with each type of evidence and explains how the various types of evidence fit together'
Good (1979, 1985)	Defines weight of evidence as the logarithm of the ratio of the likelihood of a given hypothesis to the likelihood of an alternative hypothesis. This expression corresponds to the Bayes factor
Hope and Clarkson (2014)	Refers to Good for quantitative definition Describes weight of evidence as 'basically the process of considering the strengths and weaknesses of various pieces of information in order to inform a decision being made among competing alternatives'
Hull and Swanson (2006)	Describes weight of evidence as 'approaches (that) integrate various types of data (e.g., from chemistry, bioassay, and field studies) to make an overall conclusion of risk'
Linkov et al. (2009)	Defines weight of evidence as 'a framework for synthesising individual lines of evidence, using methods that are either qualitative (examining distinguishing attributes) or quantitative (measuring aspects in terms of magnitude) to develop conclusions regarding questions concerned with the degree of impairment or risk'

Publication	Definitions or descriptions given for weight of evidence
NRC (2009)	States that 'The phrase weight of evidence is used by EPA and other scientific bodies to describe the strength of the scientific inferences that can be drawn from a given body of evidence'.
Lorenz et al. (2013)	Defines 'weight of evidence framework' as 'approaches that have been developed for taking the process from scoping an assessment and initial identification of relevant studies through the drawing of appropriate conclusions'
Schleier et al. (2015)	Describes weight of evidence as 'approaches in which multiple lines of evidence can be considered when estimating risk'
Suter and Cormier (2011)	'In sum, weighing evidence is a synthetic process that combines the information content of multiple weighted pieces of evidence. The information may be dichotomous (supports or not), quantitative values (e.g., an exposure or risk estimate), qualitative properties (e.g., large, medium or small), or a model. The weights that are applied to the information may express various properties that affect its credibility or importance and the weights themselves may be qualitative or quantitative. The combining of evidence may be a simple quantitative operation (e.g., weighted averages of concentration estimates) but more often involves difficult qualitative judgments'
Vermeire et al. (2013)	Implicit definition: 'The different and possibly contradictory information is weighted and the respective uncertainties taken into account in a weight of evidence approach'
Weed (2005)	Identifies three characteristic uses of the term weight of evidence: metaphorical, methodological and theoretical. Does not propose a definition but recommends that authors using weight of evidence should define the term and describe their methods
WHO (2009)	Defines weight of evidence as 'a process in which all of the evidence considered relevant for a risk assessment is evaluated and weighted'
SCENIHR (2012), Meek et al. (2014)	Uses the term weight of evidence but do not include an explicit definition or summary description
Rooney et al. (2014) (OHAT), Morgan et al. (2016) (GRADE)	These publications do not use the term weight of evidence but rather use related terms including 'evidence synthesis' and 'evidence integration'

Table B.2: Examples of definitions and descriptions of 'line of evidence' from the published literature

Ågerstrand and Beronius (2016)	Line of evidence used only when quoting EFSA (2013) Guidance on assessment of pesticide risks to aquatic organisms
ANSES (2016)	Defines line of evidence as 'A set of relevant information of similar type grouped to assess a hypothesis'
Bradford Hill (1965)	Not used
Collier et al. (2016)	Uses line of evidence frequently. Does not provide an explicit definition, but Table A.1 appears to imply that the evidence relating to the molecular initiating event, the adverse outcome, and to each key event is considered as one line of evidence in each case
ECHA (2010)	Line of evidence not mentioned as such. Refers to weighing pieces of evidence; also refers to weight of evidence providing 'the opportunity to make use of less reliable information/studies when they are pooled together with other information'. However, 'pooling' here may refer to the body of evidence as a whole rather than to creating subsets of evidence
EFSA (2013) PPR Panel Aquatic RA Guidance	States that 'the contribution of the multiple assessment approaches (multiple lines of evidence) in reducing overall uncertainty can ... be evaluated by weight of evidence in the final risk characterisation', implying that in this context an line of evidence is an 'assessment approach'

US EPA (1998)	Two page section on lines of evidence 'The development of lines of evidence provides both a process and a framework for reaching a conclusion regarding confidence in the risk estimate' 'Confidence in the conclusions of a risk assessment may be increased by using several lines of evidence to interpret and compare risk estimates. These lines of evidence may be derived from different sources or by different techniques relevant to adverse effects on the assessment endpoints, such as quotient estimates, modeling results, or field observational studies' 'The phrase lines of evidence is used to de-emphasise the balancing of opposing factors based on assignment of quantitative values to reach a conclusion about a 'weight' in favor of a more inclusive approach, which evaluates all available information, even evidence that may be qualitative in nature. It is important that risk assessors provide a thorough representation of all lines of evidence developed in the risk assessment rather than simply reduce their interpretation and description of the ecological effects that may result from exposure to stressors to a system of numeric calculations and results'
EPA (2003)	Refers to use of line of evidence by US EPA (1998) guideline for ecological RA, but does not use the term further
Hope and Clarkson (2014)	Line of evidence used extensively but no explicit definition
Hull and Swanson (2006)	Does not define line of evidence but refers to toxicity tests and population or community survey measures as examples of lines of evidence
Linkov et al. (2009)	Uses line of evidence several times. Does not provide own definition but when reviewing the US EPA (1998) guidance for ecological RA says that 'a weight of evidence evaluation treats each assessment and measurement endpoint as an individual line of evidence'
Meek et al. (2014)	Line of evidence not used
Morgan et al. (2016) (GRADE)	Line of evidence not used
National Research Council (U.S.) Committee on Improving Risk Analysis Approaches Used by the U.S. EPA (2009)	Not used
Lorenz et al. (2013)	Line of evidence used several times but no definition found
Rhomberg et al. (2015)	Line of evidence used in one place without definition. Also, uses 'lines of argument (or hypotheses)'
Rooney et al. (2014) (OHAT)	Line of evidence not used. Text refers to studies and body of evidence, and to 'streams of evidence' which are described as referring to the specific categories of human data, animal data, and 'other relevant data including mechanistic or *in vitro* studies'
SCENIHR (2012)	Uses 'lines of evidence' throughout. Does not provide an explicit definition but page 9 gives lists of lines of evidence, e.g. for hazard assessment the list comprises epidemiology studies, human volunteer studies, other human data, animal studies, *in vitro* studies, *in silico* studies, mathematical modelling, and mechanistic/mode of action studies; while for exposure assessment the list comprises exposure measurements, mathematical modelling and toxicokinetics
Schleier et al. 2015	Do not define line of evidence but refer to 'integrating multiple lines of evidence from different study types'
Suter and Cormier (2011)	Uses 'Categories of evidence' in similar manner to line of evidence. Does not provide an explicit definition for Categories. Includes 'lines of evidence' in list of keywords and uses it several times but does not explain how it relates to 'Categories'. Also uses 'body of evidence' which is defined in their Section 3.3.1 as all of the weighted categories of evidence for a hypothesis, but is also used in the first part of their Sections 3–10 studies of the same type, i.e. 'body' also can refer to multiple pieces of evidence in a single category
Vermeire et al. (2013)	Line of evidence not used
Weed (2005)	Line of evidence used once, when quoting a US EPA document
WHO (2009)	Line of evidence not in glossary or cumulative index

Table B.3: Examples of criteria for weighing evidence from the published literature, mapped onto the three basic concepts of reliability, relevance and consistency introduced in Section 2.5

Publication	Reliability	Relevance	Combination of reliability and relevance	Consistency	Other
Bradford Hill (1965)		Temporality Experimentation Specificity	Strength of association Biological gradient	Consistency of association Biological plausibility Coherence Analogy	
Collier et al. (2016)	Uncertainty and variability (treatment of)	Applicability and utility Essentiality of key events	Soundness Evaluation and peer review (extent of) Clarity and completeness (of reporting)	Consistency Biological concordance Concordance of empirical observations among key events Analogy (to other chemicals)	
ECHA (2010)	Reliability	Relevance		Quantity (in particular if contradictory info is present)	Adequacy for classification and RA
US EPA (1998)	Adequacy and quality of data Degree and type of uncertainty associated with the evidence	Relationship of the evidence to the risk assessment questions			
EPA (2003)	Uncertainty and variability (treatment of)	Applicability and utility	Soundness Clarity and completeness (of reporting) Evaluation and peer review (extent of)		
Hope and Clarkson (2014)	Study quality	Site specificity Spatial representativeness Temporal representativeness Specificity to stressor	Use of standard methods Endpoint/attribute association Exposure/response function Sensitivity to stressor Quantification of response		
Hull and Swanson (2006)		Specificity of cause	Magnitude Biological gradient/strength Uncertainty Spatial correlation Temporal correlation	Consistency of association Experimental verification	Plausibility: mechanism Plausibility: stressor
		Essentiality of key events			

Publication	Reliability	Relevance	Combination of reliability and relevance	Consistency	Other
Meek et al. (2014)				Consistency Biological concordance Concordance of empirical observations among key events Analogy (to other chemicals)	
Morgan et al. (2016) (GRADE)	Risk of bias Imprecision Publication bias	Indirectness Confounding Study design (randomised or observational)	Effect size Dose response	Inconsistency	
Lorenz et al. (2013)	Study design Bias/chance Reliability Statistical methods Internal consistency	Confounders Temporality Relevance	Strengths & weaknesses Dose response Predictivity Strength of association	Replicability (if observed) Biological plausibility	Adequacy
Rooney et al. (2014) (OHAT)	Risk of bias (15 subquestions) Imprecision Publication bias Rare outcomes	Indirectness Residual confounding	Effect magnitude Dose response	Consistency	'Other' (unspecified)
SCENIHR (2012)	Quality Reliability	Relevance/potential importance The characterisation of the stressor The relevance of the set of data for a particular endpoint	Utility (combining quality and relevance) Soundness and appropriateness of the methodology used The extent to which the full details of methodology are provided	The reproducibility of findings between experiments Consistency	Validity Uncertainties in the judgement used
Suter and Cormier (2011)	Performance Statistical analysis Potential for bias	Relevance Inherent weights of study types (e.g. randomised vs observational, field vs lab)	Study design Reporting Strength	Number of pieces Coherence Diversity	Case-specific criteria
Vermeire et al. (2013)	Sensitivity Reliability	Relevance Specificity	Predictivity		Adequacy Validity

References

Good IJ, 1979. "Studies in the History of Probability and Statistics. XXXVII A. M. Turing's Statistical Work in World War II." Biometrika, 66, 393–396.

Good IJ, 1985. "Weight of Evidence: A Brief Survey." Bayesian Statistics, 2, 249–270.

Gosling JP, Hart A, Owen H, David M, Li J, and MacKay C, 2013. "A Bayes linear approach to weight-of-evidence risk assessment for skin allergy." Bayesian Analysis, 8, 169–186.

Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, Alderson P, Glasziou P, Falck-Ytter Y, and Schunemann HJ, 2011. "GRADE guidelines: 2. Framing the question and deciding on important outcomes." Journal of Clinical Epidemiology, 64, 395–400. https://doi.org/10.1016/j.jclinepi.2010.09.012

Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, Devereaux PJ, Montori VM, Freyschuss B, Vist G, Jaeschke R, Williams Jr JW, Murad MH, Sinclair D, Falck-Ytter Y, Meerpohl J, Whittington C, Thorlund K, Andrews J, and Schünemann HJ, 2011. "GRADE guidelines 6. Rating the quality of evidence - imprecision." Journal of Clinical Epidemiology, 64, 1283–1293. https://doi.org/10.1016/j.jclinepi.2011.01.012

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Glasziou P, Jaeschke R, Akl EA, Norris S, Vist G, Dahm P, Shukla VK, Higgins J, Falck-Ytter Y, and Schünemann HJ, 2011. "GRADE guidelines: 7. Rating the quality of evidence - inconsistency." Journal of Clinical Epidemiology, 64, 1294–1302. https://doi.org/10.1016/j.jclinepi.2011.03.017

Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, Alonso-Coello P, Djulbegovic B, Atkins D, Falck-Ytter Y, Williams Jr JW, Meerpohl J, Norris SL, Akl EA, and Schünemann HJ, 2011. "GRADE guidelines: 5. Rating the quality of evidence - publication bias." Journal of Clinical Epidemiology, 64, 1277–1282. https://doi.org/10.1016/j.jclinepi.2011.01.011.

Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, Atkins D, Kunz R, Brozek J, Montori V, Jaeschke R, Rind D, Dahm P, Meerpohl J, Vist G, Berliner E, Norris S, Falck-Ytter Y, Murad MH, and Schünemann HJ, 2011. "GRADE guidelines: 9. Rating up the quality of evidence." Journal of Clinical Epidemiology, 64, 1311–1316. https://doi.org/10.1016/j.jclinepi.2011.06.004

Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams Jr J, Atkins D, Meerpohl J, and Schünemann HJ, 2011. "GRADE guidelines: 4. Rating the quality of evidence - study limitations (risk of bias)." Journal of Clinical Epidemiology, 64, 407–415. https://doi.org/10.1016/j.jclinepi.2010.07.017

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Falck-Ytter Y, Jaeschke R, Vist G, Akl EA, Post PN, Norris S, Meerpohl J, Shukla VK, Nasser M, and Schünemann HJ, 2011. "GRADE guidelines: 8. Rating the quality of evidence - indirectness." Journal of Clinical Epidemiology, 64, 1303–1310. https://doi.org/10.1016/j.jclinepi.2011.04.014

Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, Atkins D, Kunz R, Montori V, Jaeschke R, Rind D, Dahm P, Akl EA, Meerpohl J, Vist G, Berliner E, Norris S, Falck-Ytter Y, and Schünemann HJ, "GRADE guidelines: 11. Making an overall rating of confidence in effect estimates for a single outcome and for all outcomes." Journal of Clinical Epidemiology, 66, 151–157.

Hill AB, 1965. "The Environment and Disease: Association or Causation?" Proceedings of the Royal Society of Medicine, 58, 295–300.

Hope BK, and Clarkson JR, 2014. "A Strategy for Using Weight-of-Evidence Methods in Ecological Risk Assessments." Human and Ecological Risk Assessment, 20, 290–315. https://doi.org/10.1080/10807039.2013.781849

Klimisch HJ, Andreae M, and Tillmann U, 1997. "A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data." Regulation of Toxicology Pharmacology, 25, 1–5. https://doi.org/10.1006/rtph.1996.1076

Linkov I, Loney D, Cormier S, Satterstrom FK, and Bridges T, 2009. "Weight-of-evidence evaluation in environmental assessment: review of qualitative and quantitative approaches." Science of the Total Environment, 407, 5199–5205. https://doi.org/10.1016/j.scitotenv.2009.05.004

Meek ME, Boobis A, Cote I, Dellarco V, Fotakis G, Munn S, Seed J, and Vickers C, 2014. "New developments in the evolution and application of the WHO/IPCS framework on mode of action/species concordance analysis." Journal of Applied Toxicology, 34, 1–18.

Meek ME, Palermo CM, Bachman AN, North CM, and Jeffrey Lewis R, 2014. "Mode of action human relevance (species concordance) framework: Evolution of the Bradford Hill considerations and comparative analysis of weight of evidence." Journal of Applied Toxicology, 34, 595–606. https://doi.org/10.1002/jat.2984

OHAT, 2015. Handbook for Conducting a Literature-Based Health Assessment Using OHAT Approach for Systematic Review and Evidence Integration. Research Triangle Park, NC: OHAT.

Rhomberg LR, Goodman JE, Bailey LA, Prueitt RL, Beck NB, Bevan C, Honeycutt M, Kaminski NE, Paoli G, Pottenger LH, Scherer RW, Wise KC, and Becker RA, 2013. "A survey of frameworks for best practices in weight-of-evidence analyses." Critical Reviews in Toxicology, 43, 753–784. https://doi.org/10.3109/10408444.2013.832727

Rhomberg L, 2015. "Hypothesis-Based Weight of Evidence: An Approach to Assessing Causation and its Application to Regulatory Toxicology." Risk Analysis, 35, 1114–1124. https://doi.org/10.1111/risa.12206

Rooney AA, Boyles AL, Wolfe MS, Bucher JR, and Thayer KA, 2014. "Systematic review and evidence integration for literature-based environmental health science assessments." Environmental Health Perspectives, 122, 711–718. https://doi.org/10.1289/ehp.1307972

SCENIHR, 2012. Memorandum on the use of the scientific literature for human health risk assessment purposes – weighing of evidence and expression of uncertainty. Brussels: European Commission.

Schleier Iii JJ, Marshall LA, Davis RS, and Peterson RK, 2015. "A quantitative approach for integrating multiple lines of evidence for the evaluation of environmental health risks." Peer Journal, 3, e730. https://doi.org/10.7717/peerj.730

US EPA, 1998. Guidelines for Ecological Risk Assessment. Washington D.C: U.S. Environmental Protection Agency.

US EPA, 2003. A summary of general assessment factors for evaluating the quality of scientific and technical information. Washington, DC: U.S. Environmental Protection Agency.

Annex C – Examples of the application of approaches for assessing weight of evidence in different areas of work of EFSA

C.1. NDA Panel example

Summary of the weight of evidence assessment for the Scientific Opinion on the substantiation of a health claim related to vitamin D and risk of falling pursuant to Article 14 of Regulation (EC) No 1924/2006

Background information

Information about how the evidence is weighted in scientific assessments for health claims other than those related to well-established functions of essential nutrients can be extracted from existing guidance documents to applicants (EFSA NDA Panel, 2016a,b). For nutrition claims, and for health claims related to well-established functions of essential nutrients, the scientific evidence is generally not weighed.

The main question to be addressed is always the same. Is a health claim related to a specific food/constituent[4] and to a specific health effect scientifically substantiated? This main question can be broken down into three subquestions, namely:

1) Is the food/constituent sufficiently characterised?
2) Is the claimed effect a beneficial physiological effect (relevant to human health) for the target population and can it be measured *in vivo* in humans?)
3) Is there a cause and effect relationship between the consumption of the food/constituent and the claimed effect (for the target population, under the proposed conditions of use)?

A negative answer to any of the above-mentioned questions could stop the scientific evaluation by the NDA Panel leading to a negative conclusion (i.e. the proposed health claim is not scientifically substantiated). If a cause and effect relationship is considered to be established, an additional question should be addressed in order to establish the conditions of use for the claim:

4) What is the (lowest) effective dose and the pattern of consumption required to obtain the claimed effect?

Human studies are needed for the scientific substantiation of health claims. While animal and *in vitro* studies can be used to assess the biological plausibility of the effect, they alone cannot substantiate a health claim, and thus guidance documents contain little information on how to appraise these studies individually or weigh them within the totality of the evidence.

Each relationship between a food/constituent and a claimed effect is assessed by the NDA Panel separately on a case by case basis for specific claim applications. Pertinent human studies are an absolute requirement for the scientific substantiation of health claims, and pertinent human efficacy studies are at the top of the hierarchy that informs decisions on substantiation. However, there is no pre-established rule as to how many or which types of studies are needed for substantiation. The reproducibility of the effect of the food/constituent, as indicated by the consistency of the findings (within and across studies), and the biological plausibility of the effect are also considered.

A hierarchy of evidence for substantiation is given as follows (in decreasing order of importance):

a) Human intervention studies
b) Human observational studies: prospective cohort studies, nested-case control or case-cohort studies, cross-sectional studies, ecological studies
c) Summary studies (systematic reviews, meta-analysis).

There are a series of questions to be considered for each type of human study to decide on whether to include them or not among the totality of evidence which will be pertinent to the claim (i.e. on whether or not they will be considered in the weighing of the evidence). For human intervention studies, aspects related to their relevance, such as how the intervention relates to the food/constituent that is the subject of the health claim, how the study population relates to the target population for the claim and how the outcome variables relate to the claimed effect, are considered first. Relevant (pertinent) human intervention studies are then assessed in relation to their reliability (risk of bias) by

[4] Food/constituent means a food category, a food or a food constituent (e.g. a nutrient or other substance, or a fixed combination of nutrients/other substances)

carefully considering aspects such as randomisation, appropriateness of the control group, the use of a placebo, blinding, whether the duration of the intervention is sufficient to observe the expected changes, and whether the statistical analysis of data is appropriate. For observational studies, appropriate control for confounders is an important aspect. No scoring system is in place, however, to rate the overall risk of bias of individual studies.

If a summary publication (including systematic reviews and meta-analysis) is provided for the scientific substantiation of a claim, the Panel reviews the primary data to ensure that all the individual studies included are relevant (pertinent) to the claim. Meta-analysis can provide information about the reproducibility and consistency of the effect across studies and study groups, about the dose–response relationship, and about the minimum effective dose of the food/constituent which is required to obtain the claimed effect (i.e. to establish conditions of use). The NDA Panel, however, has not relied so far on the results of meta-analyses to make a scientific judgement on whether a cause and effect relationship between the consumption of the food/constituent and the claimed effect has been established (EFSA NDA Panel, 2016a).

Health claim related to vitamin D and risk of falling

This is a health claim pursuant to Article 14 of Regulation (EC) No 1924/2006 and falls under the scope of disease risk reduction.

In order to assess the scientific substantiation of a health claim related to vitamin D and risk of falling, the Panel considered the following in relation to the first two questions (EFSA NDA Panel, 2011):

1) The food/constituent proposed by the applicant, vitamin D (D_2 and D_3), was sufficiently characterised.
2) The claimed effect was 'reduces the risk of falling. Falling is a risk factor for fractures', and the proposed target population for the claim was men and women 60 years of age and older. Risk of falling is an established risk factor for fractures in the target population, and can be assessed in human studies as the number of falls per person per observation time (incidence), the total number of falls and/or the number of subjects falling at least once).

To complete the assessment of the scientific substantiation of the claim, two main questions remain to be assessed:

3) Is there a cause and effect relationship between the consumption of vitamin D and the risk of falling in men and women 60 years of age and older?If so,
4) What is the (lowest (4a)) dose of vitamin D and the pattern of consumption required to reduce the risk of falling in men and women 60 years and older (4b)?[5]

Tables C.1 and C.2 below summarise how the NDA Panel weighed and integrated the evidence in order to answer questions 3 and 4, respectively. Different sections of the scientific opinion (EFSA NDA Panel, 2011) are cross-referenced.

Table C.1: Summary of how the NDA Panel weighed and integrated the evidence in order to answer question 3.

Question		**Is there a cause and effect relationship between the consumption of vitamin D and the risk of falling in men and women 60 years of age and older?**
Assemble the evidence	Select evidence	The opinion is based on the evidence provided by the applicant. Details on the literature search and inclusion/exclusion criteria applied by the applicant are given in the first paragraph of Section 3 of the opinion. The Panel selected the studies/meta-analysis relevant (pertinent) to the claim by considering whether they were designed to address the specific question (see second paragraph of Section 3 of the opinion and first paragraph under 'randomized controlled trials')

[5] It should be noted that question 4 is conditional to question 3 (i.e. it is only addressed if the answer to question 3 is positive).

Question		Is there a cause and effect relationship between the consumption of vitamin D and the risk of falling in men and women 60 years of age and older?
	Lines of evidence	**LOE 1.** Six RCTs investigating the effects of vitamin D supplementation on the risk of falling in the target population **LOE 2.** Five observational studies investigating the association between vitamin D supplementation and/or vitamin D status (as a surrogate marker of total vitamin D intake) and risk of falling in the target population **LOE 3.** Data (from RCTs and observational studies) and background expert knowledge on the mechanisms by which vitamin D could reduce the risk of falling (biological plausibility of the effect)
Weigh the evidence	Methods	**LOE 1.** Narrative based on expert discussion **LOE 2.** Narrative based on expert discussion **LOE 3.** Narrative based on expert discussion
	Results	**LOE 1.** Five RCTs showed an effect of vitamin D on the risk of falling in the target population at daily doses of 800–1,000 I.U. (20–25 μg); one four-arm study using vitamin D doses of 200–800 I.U. (5–20 μg) did not show an effect, but it might have been underpowered for that outcome (see Section 3 of the opinion, penultimate paragraph under 'randomized controlled trials' **LOE 2**. Results from the observational studies provided were inconsistent; residual confounding could not be excluded (see Section 3 of the opinion under 'observational studies') **LOE 3.** Given the well-established role of vitamin D on muscle function, it is biologically plausible (but still to be established) that vitamin D supplementation could improve muscle strength, physical performance and body balance in the target population (see Section 3 of the opinion, first paragraph under 'mechanisms of action')
Integrate the evidence	Methods	**LOE 2** was dismissed, rather than integrated with **Line 1**. Integration of **LOEs 1** and **3** was done by expert discussion as explained in the last two paragraphs but one of Section 3 in the opinion
	Results	The Panel concludes that a cause and effect relationship has been established between the intake of vitamin D and a reduction in the risk of falling (Section 3 of the opinion, last paragraph) in the target population for the claim, which is men and women 60 years of age and older (Section 5 of the opinion)

Table C.2: Summary of how the NDA Panel weighed and integrated the evidence in order to answer question 4

Questions		What is the (lowest (4a)) dose of vitamin D and the pattern of consumption required to reduce the risk of falling in men and women 60 years and older (4b)?
Assemble the evidence	Select evidence	The applicant provided a meta-analysis of eight RCTs which aimed to investigate the efficacy of supplemental vitamin D with or without calcium in preventing falls among older individuals. Two of the RCTs, however, were not considered pertinent to the claim, and thus secondary analyses were conducted to assess the dose-response relationship and the risk of publication bias (see Section 3 of the opinion under 'meta-analysis of randomized controlled trials')
	Lines of evidence	**LOE 1.** Six RCTs pertinent to the claim **LOE 2.** Funnel plot for risk of publication bias analysis

Questions		What is the (lowest (4a)) dose of vitamin D and the pattern of consumption required to reduce the risk of falling in men and women 60 years and older (4b)?
Weigh the evidence	Methods	**LOE 1.** Meta-analysis of the six RCT pertinent to the claim **LOE 2.** Assessment of the risk of publication bias
	Results	**LOE 1.** No conclusions can be drawn from the meta-analysis with respect to the effect of vitamin D supplementation on the risk of falling at doses of 200–600 I.U./day (5–15 μg/day). No dose–response can be identified. The meta-analysis consistently shows, however, a significant effect of vitamin D supplementation on the risk of falling (RR = 0.83; 95% CI: 0.75–0.92) at doses of 800–1,000 I.U./day (20–25 g/day) **LOE 2.** No significant publication bias was identified. The risk of bias was quantified
Integrate the evidence	Methods	The Copas model was used to adjust the effect of vitamin D supplementation at doses of 800–1,000 I.U./day (20–25 g/day) on the risk of falling for possible publication bias (RR = 0.85; 95% CI: 0.75–0.96)
	Results	The available data do not provide information about the lowest effective dose of vitamin D needed to obtain the claimed effect (**subquestion 4a**). In order to obtain the claimed effect, 800 I.U. (20 μg) of vitamin D from all sources should be consumed daily (**subquestion 4b**) (see Section 3 of the opinion under 'meta-analysis of randomized controlled trials', last paragraph; see also Section 5 of the opinion on conditions and restrictions of use)

References

EFSA NDA panel (EFSA Panel on Dietetic Products, Nutrition and Allergies), 2011. Scientific Opinion on the substantiation of a health claim related to vitamin D and risk of falling pursuant to Article 14 of Regulation (EC) No 1924/2006. EFSA Journal 2011;9(9):2382, 18 pp. https://doi.org/10.2903/j.efsa.2011.2382

EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies), 2016a. General scientific guidance for stakeholders on health claim applications. EFSA Journal 2016;14(1):4367, 38 pp. https://doi.org/10.2903/j.efsa.2016.4367

EFSA NDA Panel (EFSA Panel on Dietetic Products, Nutrition and Allergies), 2016b. Turck D, Bresson J-L, Burlingame B, Dean T, Fairweather-Tait S, Heinonen M, Hirsch-Ernst KI, Mangelsdorf I, McArdle HJ, Naska A, Neuhäuser-Berthold M, Nowicka G, Pentieva K, Sanz Y, Sjödin A, Stern M, Tomé D, Van Loveren H, Vinceti M, Willatts P, Martin A, Strain JJ, Heng L, Valtueña Martínez S and Siani A, 2017. Scientific and technical guidance for the preparation and presentation of a health claim application (Revision 2). EFSA Journal 2017;15(1):4680, 31 pp. https://doi.org/10.2903/j.efsa.2017.4680

C.2. PPR Panel

Summary of the weight of evidence assessment for the substantiation of pesticidal active substances to be included in Cumulative Assessment Groups (CAGs)

Background

In 2013, the PPR Panel developed a methodology to identify pesticidal active substances to be included in Cumulative Assessment Groups (CAGs). It was assumed that compounds belonging to the same CAG can be treated in cumulative risk assessment as if they were simple dilution of one other and will follow the dilution principles of dose addition (EFSA, 2013).

The methodology was intended to address cumulative effects in relation to maximum residue limits (MRLs) setting. Four levels of grouping were proposed, each indicating a refinement step in the methodology.

CAG level 1: common toxicological target organ;
CAG level 2: common specific phenomenological effect;
CAG level 3: common mode of action (if available);
CAG level 4: common mechanism of action (if available).

Since information on mode and mechanism of action is frequently not available, refinement to CAG level 3 and 4 was inconclusive for most of the CAGs and the induction of the same phenomenological

effect was deemed sufficient for accepting similar action and therefore justifies dose addition. The methodology for grouping substances in CAG level 2 involves the identification and characterisation of the specific effect. Identification of the specific effect was based on the following criteria: exclusion of local effect, exclusion of non-adverse effects, exclusion of effects not relevant to humans, evaluation of unambiguous nature of the effect, identification of non-specific effect. The characterisation of the specific effect is described by supporting indicators, e.g. histological, biochemical or clinical indicators. The methodology developed by the PPR Panel was substantiated by expert knowledge (EFSA, 2013).

Following establishment of the criteria, a data collection including authorised and non-authorised substances was performed by collecting data from the dossiers submitted for the authorisation procedure and the active substances matching with the established criteria were included in the CAGs. These CAGs were defined at level 2, and often contain large numbers of substances.

Assuming that all the substances in a level 2 CAG combine by dose addition may lead to a large overestimation of risk if some or many of them do not, in fact, share the same mechanism of action. Therefore it is relevant for the risk assessors and the risk managers to consider the level of confidence or certainty that a given substance belongs to a CAG at levels 3 or 4 and consequently contributes to the cumulative risk assessment (CRA) by dose addition.

Example

The following example is proposing a weight of evidence approach using a pre-established CAG level 1 nervous system and CAG level 2 autonomic response (acute). The methodology also includes the selection of a Reference Compound which was selected on the consistency and robustness of the database (e.g. dose-related effect, consistently observed across the studies, known pesticidal mode of action, known toxic mode of action). Four lines of evidence were identified for assessing whether each substance should qualify for combining with the reference compound by dose addition, and criteria were defined for weighing each line of evidence on scales expressed as 0/+ or 0/+/++ (see below for details). The lines of evidence were then integrated by expert judgement, expressing the conclusion for each substance in terms of the probability that it qualified for dose addition, expressed on a scale of 0–100%. This can then be used in cumulative risk assessment to take account of the confidence that each substance should be included or excluded in dose addition, using probability theory, whereas there would be no such theoretical basis if the conclusion was expressed qualitatively (e.g. as the number of + scores).

This approach requires that the question addressed by the probability for each substance should be well-defined. It was considered that dose addition could be assumed for a given substance (Y) if it causes a significant effect on the autonomic nervous system (CAG level 2) and has a key event (e.g. biochemical effect) that is (a) in common with substance **X** (the reference compound) and (b) has a causal relation to the adverse outcome (AO). The approach to the weight of evidence assessment for this question is summarised in Table C.3, and results for a selection of substances are shown in Table C.4. As a final step, the elicited probability was assigned to one of the categories on a probability scale suggested in the draft EFSA Guidance on Uncertainty (see Table C.5).

Table C.3: Summary of proposed approach

Question		Does substance **Y** cause a significant effect on autonomic (CAG level 2) and have a key event (e.g. biochemical effect) that is (a) in common with substance **X** (the reference compound) and (b) has a causal relation to the AO and therefore justifies dose addition? (This question is assessed separately for each substance Y in the Level 2 CAG, except for the reference substance X)
Assemble the evidence	Select evidence	Data were collected from the authorisation dossiers and draft assessment reports (DAR) using criteria defined in EFSA, 2013. Data collection is available on EFSA, 2012 and 2015
	Lines of evidence	LOE 1: specificity of the effect/dose relationship LOE 2: clinical observation LOE 3: biochemical observation LOE 4: MoA Note that lines 1 and 2 relate to CAG level 2, whereas LOEs 3 and 4 relate to CAG level 3 or 4

Question		Does substance **Y** cause a significant effect on autonomic (CAG level 2) and have a key event (e.g. biochemical effect) that is (a) in common with substance **X** (the reference compound) and (b) has a causal relation to the AO and therefore justifies dose addition? (This question is assessed separately for each substance Y in the Level 2 CAG, except for the reference substance X)
Weigh the evidence	Methods	LOE 1: 0 No dose relationship, + Effect observed at the high dose only, ++ Effect showing a dose relationship LOE 2: 0 No, + Yes LOE 3: 0 Not observed, + biochemical read-out observed at the high dose only, ++ biochemical read-out showing a dose relationship LOE 4: 0 Not established, + presumed, ++ known
	Results	See Table 2
Integrate the evidence	Methods	Expert knowledge elicitation (EKE) was the selected methodology to integrate the evidence and express the conclusion as a probability. The methodology followed the principles of the EFSA (2014) EKE guidance, modified for application to the case in hand (a binary question) and streamlined to be practical for application to multiple substances
	Results	The conclusion was expressed as the probability that substance Y causes an autonomic effect and has a shared KE with the reference substance. This can be used in a cumulative risk model to take account of the degree of confidence that Y and X should be combined by dose addition

Table C.4: Example of preliminary results. CAG: Autonomic division, Acute

Active substance	Indicator of specific effect	NO(A)EL mg/kg bw	LO(A)EL mg/kg bw	Mode/mechanism of action	Study	Lines of evidence: Is the effect specific and therefore dose related? 0 = not specific + = only observed at high dose ++ = dose related	Lines of evidence: Is the effect defined at clinical level? 0 = No, + = Yes	Lines of evidence: Is the effect defined at biochemical level? 0 = No, + = Yes, ++ = Yes, dose related	Lines of evidence: Is the effect supported by a mechanism of action? 0 = No + = presume ++ = Yes	Lines of evidence: Expert Knowledge Elicitation	Probability scale
Oxamyl (reference compound)	Salivation, urination	0.1	0.75	**Known,** inhibition of AChE	Acute neurotoxicity rat	++	+	++	++	100%	Extremely Likely
Acetamiprid	Urination	10	30	**Known**, nicotinic acetylcholine receptor (nAChR) agonist	Acute neurotoxicity rat	++	+	0	++	50%	As likely as not
beta-Cyfluthrin	Salivation	2	10 (only observed at high dose)	**Presumed**, type II (α-cyano) pyrethroid	Acute neurotoxicity rat	+	+	0	+	33%	Unlikely
Dicofol	Lacrimation salivation	25	250	Unknown	Acute neurotoxicity rat	++	+	0	0	40%	As likely as not
Tebuconazole	Salivation	250	500 (only observed at high dose)	Unknown	Acute neurotoxicity rat	+	+	0	0	30%	Unlikely

Table C.5: Draft probability scale suggested by EFSA (2016a–c)

Probability term	Subjective probability range
Extremely likely	99–100%
Very likely	90–99%
Likely	66–90%
As likely as not	33–66%
Unlikely	10–33%
Very unlikely	1–10%
Extremely unlikely	0–1%

References

EFSA, 2013. Scientific Opinion on the identification of pesticides to be included in cumulative assessment groups on the basis of their toxicological profile, EFSA Journal 2013;11(7):3293.

EFSA, 2014. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. EFSA Journal 2014;12(6):3734.

EFSA, 2012. Grant agreement CFT/EFSA/PRAS/2012/07 awarded to the consortium ANSES/ICPS/RIVM: AS screened for neurotoxicity, liver toxicity, repro&development.

EFSA, 2015c. Grant agreement GP/EFSA/PRAS/2013/02 awarded to the consortium ANSES/ICPS/RIVM: active substances screened for effects on the nervous system, thyroid, liver, repro&development, adrenals, eye.

EFSA, 2016. Guidance on uncertainty in EFSA scientific assessment. Revised draft for internal testing. Available online: https://www.efsa.europa.eu/en/topics/topic/uncertainty-assessment

C.3. FEEDAP Panel example

Subject: Scientific Opinion on safety and efficacy of Cygro® 10G (maduramicin ammonium α) for chickens for fattening. EFSA Journal 2011;9(1):1952

Background

Additives for use in feed are substances of very different nature, ranging from trace elements (such as copper or zinc, usually in the form of salts) to viable microorganisms (silage agents or probiotics). They are categorised according to their functions and properties (technological, sensory, nutritional, zootechnical, coccidiostats and histomonostats). Moreover, some applications for feed additives do not have a specific holder and the additive is evaluated in a generic way (i.e. not linked to a concrete product). For these reasons, the assessment of feed additives is complex and varies according to the nature of the product.

The evaluation of feed additives involves its characterisation, its safety and its efficacy. The safety assessment needs to answer the following questions:

Question 1. Is the additive safe for the target animals (i.e. the animals for which the additive is intended to)?
Question 2. Is the additive safe for the consumers (of the animals fed that additive)?
Question 3. Is the additive safe for the users (such as farmers and other operators handling the product)?
Question 4. Is the additive safe for the environment?

As an example, here is discussed how the evidence was weighted and integrated to answer Question 1 for one feed additive (Cygro® 10G, maduramicin ammonium α) under the category coccidiostats and histomonostats.

Evaluation of Cygro® 10G (maduramicin ammonium α) for chickens for fattening

For the evaluation of the additive Cygro® 10G (a coccidiostat with maduramicin ammonium α as active principle) for chickens for fattening, the FEEDAP Panel weighed and integrated the evidence as summarised in the following Table.

Table C.6: Summary of proposed approach

Question		Is the additive safe for the target animals?
Assemble the evidence	Select evidence	The evidence was obtained: • from the applicant in the technical dossier linked to the application • From previous opinions of the FEEDAP Panel
	Lines of evidence	1) Tolerance studies with the additive in poultry (two studies) 2) Measurements of the minimum inhibitory concentration (MIC) of maduramicin ammonium against microbial strains (two studies) 3) Studies on the compatibility of maduramicin ammonium with the antibiotic tiamulin (one study and one previous opinion from the FEEDAP Panel)
Weigh the evidence	Methods	Assessment of the quality and validity of the studies and weighting of the evidence was performed qualitatively by best professional judgement. For this, the studies were first assessed against the requirements on the corresponding guidance documents, and then checked for their methodological design and coherence of results
	Results	Main outcomes: • One tolerance study is of limited value because of the difference between the intended and measured concentration of maduramicin ammonium in feed • In the other tolerance study, zootechnical parameters are of limited value because the low number of animals tested Main outcome: • Previous studies on interactions between ionophore anticoccidials and tiamulin are well established The provided study does not enable to establish the full compatibility between both substances Main outcome:
Integrate the evidence	Methods	Best professional judgement is the method used by the FEEDAP Panel to integrate lines of evidence of different nature, in order to achieve conclusions for complex questions in which different subjects need to be answered
	Results	The different lines of evidence enabled to conclude that: • The highest proposed dose for the additive seems close to the intolerance level but it was not possible to derive a margin of safety, due to the limitations of the tolerance studies • It is unlikely that the additive will adversely affect the overall gut microbiota • In the absence of a clearly characterised compatibility, it is not advisable to use the product with tiamulin

C.4. GMO Panel Example

EFSA GMO Panel scientific opinion on the safety of the newly expressed PjΔ6D and NcΔ15D proteins in soybean MON 87769

Background information

Genetically modified organisms (GMO) are 'regulated products' under EU legislation. In particular, Regulation (EC) No 1829/2003 (European Commission, 2003) and Implementing Regulation (EU) No 503/2013 (European Commission, 2013) require regulatory approval and a premarket safety assessment of GM products before they can be introduced into the European market. The safety assessment of foods and feeds derived from GMOs, relies on an internationally harmonised stepwise,

comparative approach, in which the GMO is compared to a non-GM counterpart with a history of safe use (Codex Alimentarius, 2009; EFSA GMO Panel, 2011). This usually entails a characterisation of the GMO at various levels of molecular organisation, including the analysis of changes caused by the genetic modification at the level of newly inserted DNA and newly expressed proteins encoded by the introduced DNA. This comparison also comprises an extensive analysis of the chemical composition and agronomic/phenotypic characteristics of the GMO vs its non-GM counterpart, taking into account the natural variability of all measured endpoints, which is estimated measuring the background range of variation observed for known commercial varieties with a history of safe use. The differences identified further guide the risk assessment and allow the identification of those aspects of the GMO needing further safety evaluation via specifically designed studies, taking into account the host organism, the type of genetic modification and the outcome of the comparative assessment. A number of topics relevant for food and feed safety commonly assessed in all dossiers submitted for regulatory pre-market safety assessment of GMOs, is as follows:

- occurrence of intended and possible unintended changes of the composition and other host characteristics;
- potential toxicity of the compound(s) introduced or whose level(s) have been altered in the host organisms by the genetic modification;
- potential allergenicity of newly expressed protein(s) given that the vast majority of the constituents responsible for allergenicity of foods are proteins;
- potential impact of the GM food or feed on human and animal nutrition (*e.g.* in case the nutritional characteristics of the GMO have deliberately been modified, such as in the case of GM soybean with changed oil composition).

Moreover, the environmental impact of the introduction of the GMO into the environment, through cultivation of a GM crop by European farmers or through spillage of imported seed must be considered. Items evaluated during the environmental risk assessment of a GM crop include, but are not limited to, change in persistence and invasiveness of the GM crops, impact on the potential horizontal gene transfer of introduced DNA to recipients (e.g. microorganisms), interaction with non-target organisms and impact of the management practices applied to the GM crop.

From a general point of view, this wide variety of data straddling that serves as basis for the assessment of the safety of a GMO calls for an application of the weight of evidence approach. Yet for GMO safety assessment, 'Weight of evidence' is explicitly referred to only for the evaluation of its potential allergenicity.

For allergenicity assessment of newly expressed proteins, the weight of evidence approach followed takes into account all of the information obtained on the newly expressed protein, since no single piece of information or experimental method yields sufficient evidence to predict allergenicity. For this assessment, all available information including tests of different nature will add to the weight of evidence (EFSA, 2006; Codex Alimentarius, 2009; EFSA GMO Panel, 2011). The information relevant for the assessment will include:

- information on the source of the transgene (i.e. the new genes introduced into the host through genetic modification), i.e. does the source have a history of allergies among patients sensitive towards it?
- bioinformatics analysis searching for similarities between amino acid sequences of the newly expressed protein and those of known allergens, based on sequence information in protein sequence repositories;
- pepsin resistance test (*in vitro* degradation study under model conditions with the gastric protease pepsin) as an indicator of the likelihood that the protein may withstands enzymatic degradation (as some but not all allergenic proteins are known to sustain degradation to pepsin);
- on a case-by-case basis, specific human sera analysis with sera from patients allergic to the source of the transgene and/or to allergenic proteins showing similarity. Finally, *in vitro* cell based assays and/or *in vivo* tests using animal models (these tests are still in an experimental stage of development for this purpose but under specific circumstances they could add to the weight of evidence).

The 'weight of evidence' assessment is used to allow the risk assessor to conclude with a sufficient degree of certainty on the likelihood of the newly expressed protein not being allergenic, taking into account the outcomes of various tests, none of which in its own right would be fully predictive of allergenicity.

As an example of the application of approaches for assessing weight of evidence in GMO, the following question is addressed: are the newly expressed PjΔ6D and NcΔ15D proteins in soybean MON 87769 safe for human and animals?

An EFSA GMO Panel opinion on soybean MON 87769 has been published in 2014 (EFSA GMO Panel, 2014).

Question		**Are the newly expressed PjΔ6D and NcΔ15D proteins in soybean MON 87769 safe for humans and animals?**
Assemble the evidence	Select evidence	The EFSA GMO Panel scientific opinion is based on the evidence provided by the applicant in the initial application as well as additional information provided by the applicant, scientific comments submitted by the Member States and relevant scientific publications. Details on the evidence considered for the safety assessment is provided in Section 5.1.2 and 5.1.4 of scientific opinion on soybean MON 87769
	Lines of evidence	A summary of the evidence provided for the newly expressed PjΔ6D and NcΔ15D proteins is as follows: **LOE 1**. Available reports on the safety of the source of the transgenes, i.e. the *PjΔ6D* and *NcΔ15D* genes originate from *Primula juliae* and *Neurospora crassa*, respectively. Furthermore, information on the PjΔ6D and NcΔ15D proteins which are desaturases naturally occurring in plants **LOE 2.** Bioinformatics analyses studies of the amino acid sequences of the newly expressed PjΔ6D and NcΔ15D proteins searching for relevant similarities with known toxins and allergens **LOE 3**. The resistance to degradation by pepsin of the newly expressed proteins studied in solutions at pH~1.2. In addition, heat denaturation of the newly expressed proteins investigated under several conditions of temperature **LOE 4**. Acute toxicity testing employing single-dose oral toxicity studies with the newly expressed PjΔ6D and NcΔ15D proteins
Weigh the evidence	Methods	**LOE 1.** Best professional judgement **LOE 2**. Best professional judgement **LOE 3.** Best professional judgement **LOE 4**. Best professional judgement
	Results	**LOE 1.** The genes originate from sources that are not considered to be common allergenic food. Some species of *Primula* are known to give rise to contact dermatitis, but not reported in *P. juliae* and is primarily due to benzoquinones and related compounds. *N. crassa* is used for food preparation in some regions of the world. In addition, the PjΔ6D and NcΔ15D proteins in soybean MON 87769 are desaturases sharing partial identity with other desaturases naturally occurring in plants used for food production **LOE 2**. Bioinformatics analyses of the amino acid sequences of the newly expressed proteins in soybean MON 87769 showed no relevant similarities with known toxins or allergens **LOE 3.** Protein degradation and denaturation studies did not raise safety concerns. No intact protein was seen within 30 seconds of incubation with pepsin at pH 1.2. Several shorter fragments were observed after different incubation times. Many of these were most likely co-purified proteins. Other fragments disappeared after 10 minutes of incubation. In addition, a significant loss of immunoreactivity was observed after 15 minutes at 95°C **LOE 4.** No adverse effects of the PjΔ6D and NcΔ15D proteins were observed at the doses tested when administered to CD-1 mice

Question		Are the newly expressed PjΔ6D and NcΔ15D proteins in soybean MON 87769 safe for humans and animals?
Integrate the evidence	Methods	**LOE 4** was of little value for the risk assessment of the repeated consumption of food and feed from GM plants by humans and animals. Integration of **LOEs 1, 2** and **3** was done by best professional judgement as explained in the paragraph below
	Results	The PjΔ6D and NcΔ15D proteins are integral membrane fatty acid desaturases and the scientific literature does not indicate that known toxic proteins have such desaturase activity as a component of their biological activity. Furthermore, these proteins or their sources are not considered common allergenic food. Bioinformatics did not reveal amino acid sequence homology of the PjΔ6D and NcΔ15D proteins with known toxins or allergens. Humans and animals consume other desaturases daily with no reported adverse effects. The PjΔ6D and NcΔ15D intact proteins were degraded by pepsin *in vitro* The EFSA GMO Panel concludes that there are no indications from the information available to suppose that the specific desaturases PjΔ6D and NcΔ15D in soybean MON 87769 would introduce safety concerns for humans and animals

References

Codex Alimentarius, 2009. Foods Derived from Modern Biotechnology. Codex Alimentarius Commission, Joint FAO/WHO Food Standards Programme, Rome, Italy, 85 pp.

European Commission, 2003. Regulation (EC) No 1829/2003 of the European Parliament and of the Council of 22 September 2003 on genetically modified food and feed. Official Journal of European Union L268, 1–23.

European Commission, 2013. Commission Implementing Regulation (EU) No 503/2013 of 3 April 2013 on applications for authorisation of genetically modified food and feed in accordance with Regulation (EC) No 1829/2003 of the European Parliament and of the Council and amending Commission Regulations (EC) No 641/2004 and (EC) No 1981/2006. Official Journal of European Union L157, 1–48.

EFSA (European Food Safety Authority), 2006. Guidance document of the Scientific Panel on Genetically Modified Organisms for the risk assessment of genetically modified plants and derived food and feed. EFSA Journal 2006, 99, 1–100.

EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms), 2011. Scientific Opinion on Guidance for risk assessment of food and feed from genetically modified plants. EFSA Journal 2011;9(5):2150, 37 pp. https://doi.org/10.2903/j.efsa.2011.2150

EFSA GMO Panel (EFSA Panel on Genetically Modified Organisms), 2014. Scientific Opinion on application (EFSA-GMO-UK-2009-76) for the placing on the market of soybean MON 87769 genetically modified to contain stearidonic acid, for food and feed uses, import and processing under Regulation (EC) No 1829/2003 from Monsanto. EFSA Journal 2014;12(5):3644, 41 pp. https://doi.org/10.2903/j.efsa.2014.3644

C.5. ANS Panel Example

Regulation (EC) No 1331/2008 establishes a common authorisation procedure for food additives, food enzymes and food flavourings.

The common procedure lays down the arrangements for updating the lists of substances authorised to be placed in the European Union market.

According to this procedure, EFSA is requested by the European Commission to provide scientific opinions on the proposed updates of the existing positive lists. This applies to:

- adding new substances to the existing lists;
- adding, removing or changing conditions, specifications or restrictions associated with the presence of a substance on the Community list.

The information provided in the opinion of EFSA should be sufficient to ascertain whether the proposed use of the substance is safe for consumers. This includes conclusions on the toxicity of the substance, where appropriate, and possible establishment of an acceptable daily intake (ADI) expressed in a numerical form with details of a dietary exposure assessment for all food categories, including exposure of vulnerable consumer groups.

In the case of evaluation of food additives, the EFSA Panel on Food Additives and Nutrient Sources added to Food (ANS Panel) has developed specific guidance describing the scientific data required for the evaluation of a substance proposed for use as a new food additive, as well as the risk assessment paradigm used by the Panel in drawing its conclusions (EFSA ANS Panel, 2012).

The concept developed by the ANS Panel, and described in its 2012 guidance document, is a tiered approach which balances data requirements against other considerations such as use and animal welfare.

Using this tiered approach, a minimal data set applicable to all compounds has been developed under Tier 1. Compounds which are systemically absorbed or for which toxic or genotoxic effects are found in Tier 1 will require Tier 2 testing to generate more extensive data. Tier 3 defines detailed testing for specific a case-by-case basis. A diagram of the tiered approach is given in Figure C.1.

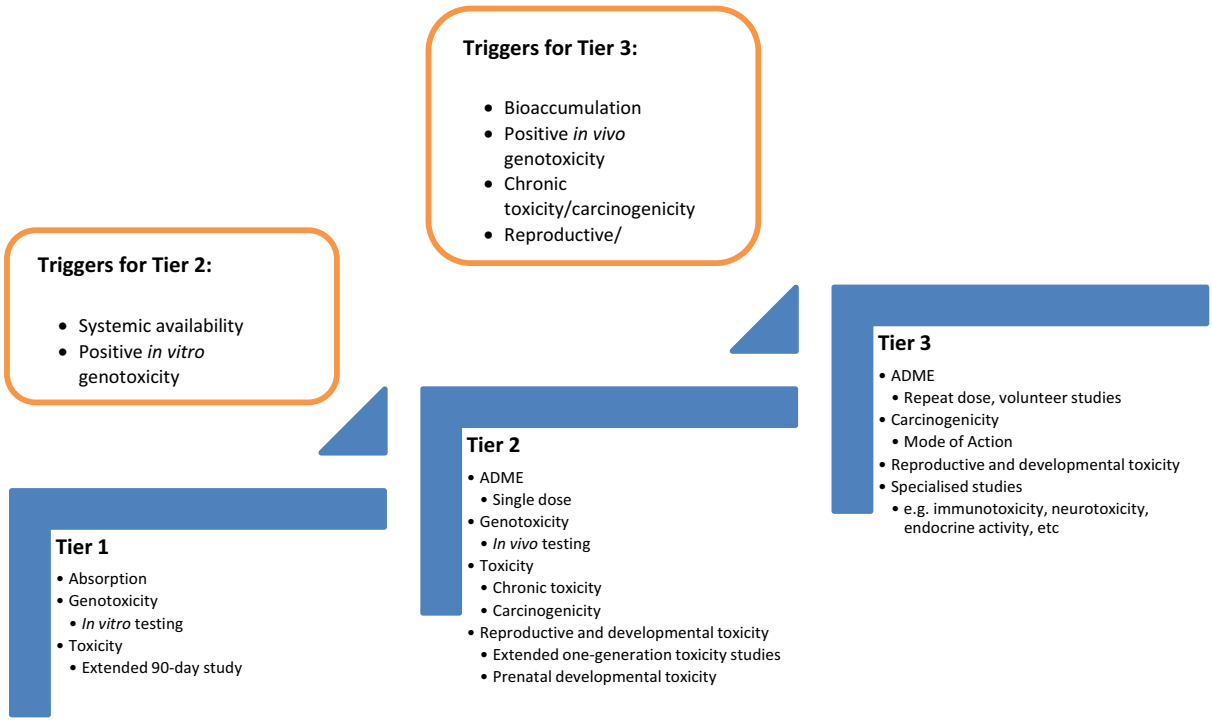

Figure C.1: Tiered approach to toxicity testing according to EFSA ANS Panel Guidance for submission for food additive evaluations (EFSA ANS Panel, 2012)

In 2016, the ANS Panel adopted a scientific opinion on the evaluation of the safety of the new food additive potassium polyaspartate (A-5D K/SD) proposed for use as a stabiliser in wine at the maximum level of use of 300 mg/L.

The opinion was further to a request from the European Commission in accordance with Regulation (EC) No 1331/2008 establishing a common authorisation procedure for food additives, food enzymes and food flavourings.

The main question to be answered by the ANS Panel was whether potassium polyaspartate (A-5D K/SD) was safe at the proposed uses for the EU population.

The toxicological data submitted by the applicant complied with the requirements for Tier 1 toxicity testing as described in the 2012 ANS Panel guidance.

Hence, the first question to be addressed by the Panel in performing the assessment was whether, on the basis of the data submitted, any triggers for Tier 2 toxicity testing could be identified for potassium polyaspartate (A-5D K/SD).

Provided that the toxicological data were considered sufficient, the Panel addressed the question on the estimated exposure to potassium polyaspartate (A-5D K/SD) from the proposed use as a stabiliser in wine at the maximum proposed level of 300 mg/L.

Only the first question will be addressed in this case study.

Based on the toxicological data provided, in line with the requirements of Tier 1 toxicity testing, the Panel considered that no further toxicity testing was necessary. The data submitted were deemed sufficient to demonstrate that proteolytic digestion of A-5D K/SD was minimal and that no absorption

of intact A-5D K/SD was observed *in vitro*. Both required *in vitro* genotoxicity tests were negative, whereas in the 90-day toxicity study in rats no adverse effects were noted at the highest dose tested of 1,000 mg/kg bw per day.

At the proposed maximum level of 300 mg/L, the mean dietary exposure to potassium polyaspartate (A-5D K/SD) ranged from 0.02 to 0.4 mg/kg bw per day in adults up to 0.05 to 0.6 mg/kg bw per day in the elderly. The high-level intake ranged from 0 to 1.4 in adults and from 0.4 to 1.8 mg/kg bw per day in the elderly. In consideration of the proposed use of potassium polyaspartate (A-5D K/SD) as a food additive, limited to wine, the Panel considered it appropriate to consider dietary exposure only in adults and in the elderly.

Based on the NOAEL of the 90-day study and these exposure estimates, the Panel considered that there would be an adequate margin of safety from the proposed use and use levels (~ 550 for the high-level elderly consumers at the proposed maximum level of 300 mg/L). The Panel noted that the NOAEL of 1,000 mg/kg bw per day was the highest dose tested and that the exposure used for this comparison is a conservative estimate because the applicant indicated that, for most wines, typical use levels of 100–200 mg/L of potassium polyaspartate would be sufficient to achieve the technological need. The Panel considers that, because of the possible uncertainties, the margin of safety estimated above is likely to be lower than the actual margin of safety.

The Panel concluded that there was no safety concern from the proposed use and use levels of potassium polyaspartate (A-5D K/SD) as a stabiliser in wine.

Question		***On the basis of the toxicological data submitted by the applicant as part of the dossier, are there any triggers for Tier 2 toxicity testing?***
Assemble the evidence	Select evidence	*The evidence was selected among the data contained in a dossier provided in support of an application for authorisation of a new food additive. The dossier was prepared in accordance with the data requirements of Tier 1 toxicity testing, described in the 2012 ANS Panel 'Guidance for submission for food additive evaluations'*
	Lines of evidence	***No 1: Systemic availability of the proposed new food additive potassium polyaspartate*** • *Sequential proteolytic attack with pepsin (porcine) and pancreatin (porcine)* • *Caco-2 cell absorption model at concentrations of 1 mg/mL* in vitro
		No 2:* In vitro *genotoxicity of the proposed new food additive potassium polyaspartate • *Bacterial reverse mutation assay performed in accordance with OECD TG 471* • In vitro *mammalian cell micronucleus assay performed in accordance with OECD TG 487*
		No 3: Subchronic oral toxicity of the proposed new food additive potassium polyaspartate • *A 14-day range-finding study performed to collect information of target organs and to select appropriate doses for a 90-day study.* • *A 90-day subchronic toxicity study (OECD TG 408), modified to include assessment of additional parameters described in the more recent guideline on repeated-dose 28-day oral toxicity study in rodents (OECD TG 407) to allow for the identification of chemicals with the potential to cause neurotoxic, immunological or reproductive organ effects or endocrine-mediated effects*

Question		***On the basis of the toxicological data submitted by the applicant as part of the dossier, are there any triggers for Tier 2 toxicity testing?***
Weigh the evidence	Methods	*The data were considered to fulfil the requirements of Tier 1 toxicity testing, described in the 2012 ANS Panel 'Guidance for submission for food additive evaluations'* *The genotoxicity and subchronic toxicity studies were standard regulatory studies carried out in recognised testing facilities according to the relevant guideline and GLP compliance and were reported in accordance with the relevant guideline (Best professional judgement)*
	Results	***No 1: Systemic availability of the proposed new food additive potassium polyaspartate*** • *Proteolytic digestion of A-5D K/SD was minimal* • *No absorption of intact A-5D K/SD was observed* in vitro ***No 2:* In vitro *genotoxicity of the proposed new food additive potassium polyaspartate*** • *Both required* in vitro *genotoxicity tests were negative* ***No 3: Subchronic oral toxicity of the proposed new food additive potassium polyaspartate*** • *The NOAEL in the 90-day toxicity study was 1,000 mg/kg bw per day, the highest dose tested, and there were no triggers for additional toxicological testing*
Integrate the evidence	Methods	*Data integrated according to the Tiered approach described in the 2012 ANS Panel 'Guidance for submission for food additive evaluations' (see Figure C.1) (Best professional judgement)*
	Results	*In view of the Tier 1 results, the Panel considered that no Tier 2 or Tier 3 testing was necessary*